Vol. 49, No. 2 March, 1952

Psychological Bulletin

CONTENTS

PUBLISHED BI-MONTHLY BY

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

The Psychological Bulletin contains evaluative reviews of the literature in various fields of psychology, methodological articles, critical notes, and book reviews. *This* JOURNAL *does not publish reports of original research or original theoretical articles.*

Editorial communications, manuscripts and book reviews should be sent to Lyle H. Lanier, Department of Psychology, University of Illinois, Urbana, Illinois.

Preparation of articles for publication. Authors are strongly advised to follow the general directions given in the article by Anderson and Valentine, "The preparation of articles for publication in the journals of the American Psychological Association" (*Psychological Bulletin*, 1944, 41, 345–376). Special attention should be given to the section on the preparation of the bibliography (pp. 363–372), since this is a particular source of difficulty in long reviews of research literature. *All copy must be double-spaced, including the bibliography.*

Reprints. Authors may order reprints when returning proof. Five copies of the JOURNAL are supplied gratis to contributors of articles, notes, special reviews, and book reviews. No reprints are supplied gratis.

Business communications—including subscriptions, orders of back issues and changes of address—should be sent to the American Psychological Association, 1515 Massachusetts Avenue, N. W., Washington 5, D. C.

Annual subscription: $7.00 (Foreign $7.50). Single copies, $1.25.

PUBLISHED BI-MONTHLY BY

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

1515 Massachusetts Ave., N.W., Washington 5, D.C.

Entered as second class mail matter at the post office at Washington, D.C., under the act of March 3, 1879. Additional entry at the post office at Menasha, Wisconsin. Acceptance for mailing at special rate of postage provided for in Section 538, act of February 28, 1925, authorized August 6, 1947. Printed in U.S.A.

Vol. 49, No. 2 March, 1952

Psychological Bulletin

ITEM ANALYSIS IN RELATION TO EDUCATIONAL AND PSYCHOLOGICAL TESTING[1]

FREDERICK B. DAVIS

Hunter College

The construction of valid and reliable tests requires consideration of quantitative information regarding the difficulty and discriminating power of each test exercise, or item, that is proposed for use. Such information is provided by item-analysis data. The development of procedures for obtaining item-analysis data and the formulation of principles governing their use are outlined and discussed in this article.

Item-Difficulty Data

The need for expressing the difficulty of a test item became evident as soon as systematic efforts to measure mental skills were undertaken. In 1905, for example, Alfred Binet, criticizing an examination constructed by Drs. Blin and Dumaye, wrote in part: "For each topic the same mark is given, thus making each topic of equal weight. One is, therefore, assuming that all topics present the same amount of difficulty" (7). Binet realized that tasks presented very different degrees of difficulty to children. By 1911, he had accumulated considerable data pertaining to the difficulty of items in his scale for measuring intelligence. In that year he noted with respect to one of the performance items in his scale: "Children of four years do not succeed in forming the oblong. Only one-third are successful. . . . By five years there is decided progress; we found that scarcely one child in twelve failed" (6).

Binet's technique for expressing difficulty as the percentage of a defined group passing an item was widely used in the years following publication of the Binet-Simon scales. Some test constructors, wishing to express item difficulty in terms of a scale having more nearly the

[1] A revision of a paper read at the session on Educational Test Theory at the Annual Meeting of the American Statistical Association, Chicago, December 27, 1950. The literature up to July 1951 is covered in this article.

desirable properties of an interval scale, proposed converting each "per cent passing" value into a corresponding standard score. So far as the writer is aware, L. P. Ayres was the first to make use of this type of item-difficulty value. He did so in constructing his well-known *Measuring Scale for Ability in Spelling* (4). At about the same time, Clifford Woody (82) made use of a similar technique in the preparation of his arithmetic scales.

In 1927, Thorndike published the results of a study of the use of "per cent passing" as an index of item difficulty (68, chap. 4). He pointed out that the level of difficulty of an item as a measure of a criterion variable not exactly identical with the variable actually measured by the item is a function of the correlation between passing or failing the item and the criterion variable as well as of the per cent passing the item in a defined group. That is, an item not highly correlated with the criterion variable and passed by only a small percentage of the group may be of the same difficulty level with respect to the criterion as another item more closely correlated with the criterion variable but passed by a larger percentage of the group. Bliss (8) dealt with this same point in 1929.

It seems likely that Thorndike's procedure for assigning a scale value to each item to indicate its difficulty level with respect to the criterion variable would be of greatest value if one were constructing a predictor test intended to provide maximum differential prediction at a designated level of ability. Most self-defining tests, such as tests of achievement, are so constructed as to measure a weighted combination of skills rather than a homogeneous mental function.

A significant advance in the assignment of difficulty indices to test items was made as early as 1927 by L. L. Thurstone. Using data pertaining to percentages passing an item in overlapping distributions, he developed absolute scaling methods (69, 70). These would undoubtedly have been used more widely than they have been if the labor involved in computing them were smaller or the need for more precise indices of item difficulty were more keenly felt.

In 1928, Brolyer (10) described the method of computing difficulty indices for the examinations of the College Entrance Examination Board. Brolyer's index differs only in its metric from one used as early as 1915 by Ayres,[2] but its method of computation takes account of the fact that not all examinees may have time to consider every item in a speeded test. Despite the fact that tryout tests are properly ad-

[2] For a definition in terms of raw scores, see (59, p. 301).

ministered with generous time limits, it is often uneconomical or impractical to allow sufficient time for every examinee to reach every item (18, pp. 278–279). If we compute the percentage of successes on an item by dividing the number who marked the item correctly by the total number of examinees, we are tacitly assuming that those examinees who did not reach the item would have done only as well as chance would have permitted had they been granted time in which to finish. If in computing the percentage we use as a denominator the number of examinees who reached the item, we are tacitly assuming that those who did not reach the item would, if granted additional time, have marked the correct answer in the same proportion as those who did answer it in the time limit. In practice, this is the more reasonable of the two assumptions, though an empirical check will usually show the percentage of correct answers to be somewhat smaller among those who did not reach the item than among those who did. This is expected since there is ordinarily a positive correlation between proficiency and speed in most mental functions (59, p. 296). Practical procedures for estimating the number of examinees not reading each item have been outlined by Davis (17, p. 33).

Brolyer's procedure also takes account of the fact that examinees may consider an item and deliberately refrain from marking an answer to it. If these individuals are not included in the denominator of the fraction used to compute the percentage of successes, some items would appear to be much easier than they really are. Although adjustments for the fact that examinees do not reach every item and do not mark an answer to every item that they reach have a good deal to recommend them, and were inherent in Brolyer's work as early as 1928, they are still not used by some test constructors.

As early as 1934, Votaw (77) suggested that item-analysis data pertaining to multiple-choice items should be corrected for chance success. In 1936, Guilford (35) discussed this matter and presented the following formula for correcting item data for chance success:

$$ {}_cP_R = P_R - \frac{P_W}{k_i - 1}, \qquad [1] $$

where ${}_cP_R$ equals the per cent of the sample that *knows* the correct answer to an item, P_R equals the per cent of the sample that *marks* the correct answer to an item, P_W equals the per cent of the sample that *marks* an incorrect answer to an item, and k_i equals the number of choices in an item.

It can be shown that this formula yields a maximum-likelihood

estimate of the per cent of a sample that actually *knows* the answer to an item provided that every examinee who does not possess enough knowledge to select the correct answer guesses among all of the choices in the item and provided that success in this guessing is governed by the binomial law (57). In actual practice, the first of these two assumptions is certainly not satisfied. Examinees often possess misinformation that leads them to mark an incorrect answer in the belief that it is correct; likewise, they often possess partially correct knowledge that is not adequate to permit them to identify the correct answer but does permit them to exclude one or more choices as incorrect. Furthermore, nonchance considerations (such as lengths of choices, etc.) that have no relation to the variable being measured lead some examinees to mark a given choice as correct or to exclude one or more choices as incorrect.

Horst (45) had thought through this problem as early as 1934 and devised a formula that took account of partial knowledge in correcting for chance success. Since the effect of nonchance considerations that have no relation to the variable being measured can be minimized by careful test construction, Horst's formula leaves only misinformation as an influential uncontrolled factor in correcting scores from well-constructed tests for chance success. Consequently, it will systematically overcorrect for chance success whenever misinformation plays a part in determining examinees' responses. When formula [1] or its variants are employed, the use of misinformation by examinees leads to *overcorrection* while the use of partial information by examinees leads to *undercorrection* for chance success. These two biasing influences tend to cancel each other. Therefore, if per cent of success for an item is to be corrected for chance success, the writer recommends the following formula:

$$_cP_R = 100 \cdot \frac{R - \dfrac{W}{k-1}}{N - NR}, \qquad [2]$$

where R equals the number of examinees who answer an item correctly, W equals the number of examinees who answer an item incorrectly, k equals the number of choices in an item, N equals the number of examinees, and NR equals the number of examinees who do not reach an item in the time limit.

The writer believes that it ordinarily does not matter greatly whether item difficulty indices are corrected for chance success, since the selection of items for a test is likely to be essentially the same whether corrected or uncorrected indices are used. However, the writer does

strongly advocate correcting scores derived from a multiple-choice test that are to be used as a criterion for item-analysis purposes. This can do no harm and is likely to provide item-analysis data that are more meaningful than would otherwise be obtained (11). Other considerations of a nonstatistical nature also lead the writer to recommend the use of criterion scores that have been corrected for chance success with the conventional formula[3] even though the assumptions underlying the use of the formula are rarely, if ever, completely fulfilled. Formulas developed by Calandra and Hamilton demand the use of practices or assumptions that are equally, or even more, inappropriate (12; 43; 66, chap. 12).

In recent years, short cut methods of determining item discriminating power have led to the use of estimates of the percentages of success derived from data pertaining to the responses of examinees in only the highest and lowest 27 per cent of the criterion-score distribution. Davis (17, p. 6) has described the method for making these estimates and has pointed out that they will be biased systematically unless the regression of item score on criterion score in the population from which any specific sample of examinees is drawn is rectilinear. Empirical studies to determine the differences between estimates of item difficulty obtained by means of the procedure described by Davis were made at the Cooperative Test Service and at the Educational Records Bureau several years ago but were not published, probably because they lack generality; that is, the amount of bias resulting from curvilinearity of regression of item score on criterion score varies from test to test. Davis (19) did publish data showing the reliability and standard error of measurement of estimates of item difficulty obtained by means of the procedure mentioned above when the upper and lower groups each comprise one hundred examinees. The reliability coefficient was so high (.98) that it is evident that any loss in reliability suffered by the omission of the middle 46 per cent of the criterion distribution is of no practical importance in test construction. Studies by Carter (13) and Gibbons (32) yielded results similar to those obtained by Davis.

The Use of Item-Difficulty Data

Maximizing test reliability. Symonds (67) and Gulliksen (39) have considered the problem of selecting items according to difficulty with the objective of maximizing the reliability of the resulting test. Their findings are in agreement and may be summarized in this quotation from Gulliksen: "If we ignore the effect of chance success, and assume

[3]For a discussion of this matter, see (18, p. 277) and (62).

that we are dealing with a 'well-constructed' test, then in order to maximize the reliability and variance of the test, items should have high intercorrelations, all items should be of the same difficulty level, and this level should be as near to 50 per cent as possible." Whether the objective of maximizing over-all test reliability is ever worth while is debatable. It can be shown that the selection of items all of 50 per cent difficulty will under some circumstances tend to lower a test's correlation with an external criterion or with true scores in the mental function measured by the entire test. It, therefore, seems doubtful to the writer that one would ever want to select items solely, or even principally, to maximize over-all test reliability.

Maximizing over-all test validity. Several investigators have studied the problem of selecting items according to difficulty in such a way as to maximize test validity. Cleeton (15) and Thurstone (71) both found that items close to 50 per cent difficulty seem to be required for this purpose. Cleeton's results actually apply only to the validity of items taken individually and have no direct bearing on the validity of tests composed of more than one item. Thelma G. Thurstone presented the correlations between scores on a 993–item spelling test and scores on each of twenty-two 25-item spelling tests included in the 993–item test. Her data indicated that the correlations were highest when the subtests included items ranging in difficulty from the 45 per cent level to the 54 per cent level. These findings led some test constructors to the incorrect generalization that maximum validity for any test would be obtained by using only items of 50 per cent difficulty. Actually, it is now known that the selection of items according to difficulty in order to maximize over-all test validity is a function of the magnitude of the item intercorrelations. As it happened, the average intercorrelation of Mrs. Thurstone's spelling items was such that a concentration of item difficulties between the 45 and the 54 per cent levels provided close to maximum possible correlation with total scores on the 993-item test. Mrs. Thurstone, however, did not generalize beyond her data and it is now clear that her findings should form a basis for selecting items only if one wants to maximize *over-all* test validity and only then when the average item intercorrelation is about the same as that of the spelling items she used.

In 1939, Flanagan (27) sketched the rationale for selecting items to maximize the sum of the discriminations that they would make among the individuals tested. Ferguson (23) has dealt with the problem of selecting items in such a way as to maximize the number of discriminations that the total scores on a test can make among individuals tested.

He has shown that this will occur if the distribution of test scores is rectangular and extends across the entire range of possible scores. Davis started from this point and demonstrated that a rectangular distribution of raw test scores can be produced, when the distribution of item difficulties is symmetrical around the 50 per cent level, by selecting items at certain specified difficulty levels that are fixed by the intercorrelations of the items.[4] If the tetrachoric item intercorrelations are all unity, a rectangular distribution of raw scores is most likely to be obtained by selecting items with difficulty levels of

$$\frac{1}{n+1}, \frac{2}{n+1}, \frac{3}{n+1}, \cdots, \frac{n}{n+1},$$

where n equals the number of items in the test. If the tetrachoric intercorrelations are all .500, a rectangular distribution of raw scores is most likely to be obtained by selecting items all of 50 per cent difficulty.

If the tetrachoric intercorrelations of the items in a test were all below .500, the total scores derived from a test in which all items were of 50 per cent difficulty would be more likely to be platykurtic than to be completely rectangular. If the item intercorrelations were all zero, it would be exceedingly unlikely that a rectangular distribution of test scores would ever occur by chance in a sample like the one in which the item difficulties were obtained. Nonetheless, for any level of tetrachoric item intercorrelations from zero to .500, the maximum number of discriminations that can be made by the total score will be secured by selecting items all at the 50 per cent level. It is thus obvious that Flanagan's and Ferguson's objectives will accomplish the same result in terms of test efficiency.

A simple mathematical procedure employed by Davis to specify the exact difficulty levels of items for 2- and 3-item tests that will be most likely to cause the resulting test scores to have maximum possible discriminating power cannot be applied to specifying the exact difficulty levels of items for tests containing larger numbers of items except in the limiting case when the item intercorrelations are all unity. But it does permit inferences to be drawn with regard to distributions of item difficulty desired in such tests, and it points to the need for a general solution for the problem of determining the exact item-difficulty levels required to maximize the likelihood that test scores will provide as much discrimination as possible.

It is apparent that item intercorrelations are kept to relatively low

[4] See (20). A revision of this paper was read at a meeting of the American Educational Research Association in Atlantic City in February, 1950.

levels because of the inevitable lack of reliability of single items. Thus Davis' findings suggest that item-difficulty indices often ought to be clustered close to the 50 per cent level in order to produce a test having maximum discrimination among all the examinees. Some years ago, Lawley (**53**) presented a rigorous development showing the distribution of item difficulties required for a test in which all items measured only one mental function plus chance, if the test scores were to provide maximum discrimination. He assumed that the one mental function measured was distributed normally in the population tested, and his results indicated that items of low reliability should be clustered near the 50 per cent level. This agrees with Davis' findings since items of low reliability would necessarily have low raw intercorrelations even though their intercorrelations after correction for attenuation were unity.

Tucker and Brogden have both studied the problem of selecting items according to difficulty in such a way as to maximize the correlation between scores derived from a test and a criterion variable, which may be either true scores on the test itself or an external variable. Tucker (**73**) deals with the special case of a test in which every item is of equal difficulty and measures with reliability equal to that of every other item only a perfectly reliable criterion variable. If we look at Tucker's Figure 4, we see that the maximum validity coefficient is obtained for a certain test of 100 items, all of 50 per cent difficulty, when the item intercorrelations are about .25. As they increase above this value, the test validity coefficient decreases until, with perfectly correlated items, it is only .798. These data show that the validity coefficient of a test made up of equivalent items will decrease as the level of item intercorrelation is increased above a certain point. This drop in validity occurs despite an accompanying increase in test reliability. The shape of this curve is influenced by the fact that item validity must change as item reliability changes. In this respect the conditions are unlike those postulated by Davis in his development. He dealt with the problem of selecting items according to difficulty only and assumed various degrees of intercorrelation for items having a fixed level of validity. Tucker's results indicate that if items of 50 per cent difficulty are used exclusively in a test (in order to maximize its reliability), its over-all validity coefficient will be at a maximum only at a particular level of item intercorrelation. This finding shows that the methods recommended by Gulliksen (**39**, p. 90) for maximizing over-all reliability are not necessarily appropriate for use in test construction if the goal is to maximize over-all test validity, as is often

the case. Tucker described the 100 items in the test for which validity coefficients were graphed in his Figure 4 as "perfectly pitched in difficulty" (73, p. 11). His data show that equivalent items of 50 per cent difficulty will produce a test having maximum validity at a rather low level of item intercorrelation. As the item reliabilities (and thus their intercorrelations and validity coefficients) are increased above a certain low level, the validity of the test decreases because the items are *imperfectly* pitched in difficulty to secure maximum validity under the changed conditions.

Brogden has approached this matter from a semi-empirical point of view and has noted this same phenomenon, which he calls "this seemingly paradoxical finding," that when test items are scored 1 for "correct" and 0 for "wrong," increasing test reliability does not necessarily increase test validity (9, p. 206). It appears from Brogden's and Davis' findings that constructing a test to provide maximum discrimination among all of the examinees is synonymous with constructing a test to have maximum over-all validity. This point needs analytic treatment.

It is *not* proper to deduce from Tucker's data that to obtain high test validity one should make use of items of low reliability. There is no inconsistency between high item reliability and efficient measurement.

Maximizing differential validity. In 1936, Richardson (64) dealt with the problem of maximizing the validity of a test used only to dichotomize a sample at a specified level of ability. Tests are often used for such a purpose with no attention being paid to differences between scores of individuals in either of the two groups thus formed. Richardson showed that maximum differential validity will be secured by constructing a test with all items at the difficulty level represented by the dichotomic line.

If we approach this problem from the point of view of discriminations among the individuals tested, it is evident that the only discriminations that are useful are those that differentiate the members of the two dichotomized groups. The number of useful discriminations would be at a maximum if every member of one group got zero scores and every member of the other group got perfect scores. This could happen if all items were of equal difficulty and had tetrachoric intercorrelations of unity. Naturally, we do not ever have items with reliability coefficients high enough to permit tetrachoric intercorrelations of unity, so not all the discriminations made by a test are useful for differentiating between individuals in the dichotomized groups even when all items are

at a single level of ability. Instead, the number of useful discriminations is merely maximized, given certain item reliability and intercorrelation. We can see that the problem of maximizing differential validity for a test is the same as that of maximizing the number of useful discriminations it can make between individuals above and below a specified level of ability. The solution is simply to place all items at the level of ability designated as the point of dichotomy regardless of their level of intercorrelation. The amount of differential validity achieved, however, is dependent on the individual item validities, reliabilities, and intercorrelations. Naturally, as item validities increase, other things being equal, the intercorrelations of the items will tend to increase; but there is a great deal of latitude at the levels of validity coefficients we ordinarily obtain.

Controlling the standard error of measurement. The purpose for which test scores are to be used should be taken into account when a test is constructed so that items can be selected according to difficulty in such a way as to control the standard error of measurement at different points on the ability scale. Although Jackson and Ferguson (49) pointed this out as long ago as 1943, very little systematic work has been done to show analytically the relationship between the shape of the distribution of item difficulty indices and the size of the standard error of measurement at various levels of ability. Lawley (53) studied the problem from a mathematical point of view and Mollenkopf (60) presented empirical data that are relevant to it. Lawley (53, p. 281) indicates that to obtain maximum discrimination at a given point on the scale of ability, the items in a test should be selected so that they are all of a level of difficulty corresponding to the given point on the scale of difficulty, whereas on page 280 he states, "It appears, therefore, that the order in which individuals are placed by their test performance is more reliable at either end of the scale than in the middle." To the writer, these statements seem contradictory because, logically, it would seem that the point at which maximum discrimination is made should coincide with the point at which accuracy of measurement is greatest. Mollenkopf's data show that if item difficulties in a given test are concentrated near the 50 per cent level, the standard errors of measurement at several raw-score levels will be larger near the middle scores than at the ends of the range of scores. If Mollenkopf's data were reworked to express scores derived from his various tests as approximations to interval scores on a single scale, the writer believes they might show that the size of the standard error of measurement in a given range of ability decreases with an increase in the number of discriminations

among examinees obtaining scores in that range of ability. Thus, if all items are concentrated at a single ability level, the minimum standard error of measurement (expressed in interval scores) would presumably be found at that ability level. Lawley's conclusion that a test places individuals in order more accurately at the ends of the scale than in the middle may result from the fact that he was dealing with raw scores rather than interval scores on a single scale of ability.

Our present state of knowledge with respect to controlling the magnitude of the standard error of measurement at various ability levels may be summarized in the following statements:

1. To minimize the aggregate of errors of measurement of a test (thus perhaps sacrificing over-all validity or differential validity), all items should be of 50 per cent difficulty. This will maximize the over-all test reliability coefficient and minimize the standard error of measurement at the center of the range of ability measured.
2. To minimize the standard error of measurement at any one point on a scale of ability, all items should be concentrated at that level of difficulty.
3. To minimize the standard error of measurement at two or more points on a scale of ability, items should be apportioned to each of the levels of difficulty specified.
4. To minimize the standard error of measurement throughout a certain part of the range of ability, items should be distributed within the corresponding range of difficulty in accordance with the procedure suggested for obtaining maximum over-all test validity.
5. To equalize as nearly as possible the standard error of measurement throughout the range of ability measured, items should be distributed over the entire range of difficulty in accordance with the procedure suggested for obtaining maximum over-all test validity.

Some readers may be surprised at the amount of space devoted to a discussion of item-difficulty indices and their uses. In the writer's opinion, the importance of these data has often been overlooked. He believes that altogether too much relative emphasis has been placed on the computation and use of indices of item discrimination. Too often, a great deal of time has been expended in trying to determine the correlation of test items with a crude criterion of doubtful value while very little time has been devoted to maximizing test efficiency by following item rationales, writing and editing items with great care, and selecting items according to difficulty.

Item-Discrimination Power

The first systematic validation of test items can probably be credited to Alfred Binet, whose procedure was simply to note the percentage of a sampling of children at successive age levels who could pass an item

in the early versions of his scales for measuring intelligence. As early as 1901 Wissler (81) correlated scores derived from certain reaction-time, memory, and color-naming tests with grades obtained by Columbia College freshmen. However, most early item analyses involved comparing the percentage passing each item in known groups, such as grade groups in school.

Graphic methods. Later it became customary to plot these data graphically and to select items on the basis of the slopes of the lines connecting the points representing percentage passed. In the absence of satisfactory criterion data, test constructors came to dividing the distribution of scores derived from a tryout test into several parts, such as thirds or fourths, and graphing the percentage of successes in each part (63, 75). A refinement of this procedure has been described by Turnbull (74). The criterion group is divided into sixths on the basis of some criterion (such as total score on the test) and the percentage of the examinees in each sixth who marked each item correctly is computed. These data for each item are then plotted on specially prepared graph paper. The percentages in each sixth that marked the item incorrectly may also be plotted on each sheet of graph paper. This procedure provides revealing data pertaining to each item but is rarely, if ever, used because practical circumstances ordinarily do not permit use of all the data obtained at the cost of a great deal of labor.

Statistical significance. Because there is no convenient method for determining whether the slope of the line on a graph showing the percentage passing an item in each one of several groups differs significantly from the horizontal, many test constructors favor computing percentages of examinees passing an item in only two groups and obtaining a critical ratio to show the statistical significance of the difference between the two percentages. This is easily accomplished by well-known procedures. If the two groups in which the percentages are computed consist of examinees having "high" and "low" scores in the criterion variable, it is best to use the highest and lowest 27 per cents of the criterion-score distribution, as shown by Kelley as early as 1928.[5]

Chi square has been proposed as a measure of item discrimination by Guilford (36) and others. To avoid some of the limitations of the critical ratio and chi square and their interpretations, Cureton has suggested the use of chi.[6] It is properly applicable to much smaller

[5] For Kelley's rationale, see (52).

[6] For computing formulas and abridged tables originally prepared by E. E. Cureton for use with the chi test, see (18, pp. 289–290). Footnotes *a* and *b* to Table 6 in (18) should be interchanged, as pointed out to the writer by J. C. Stanley.

samples than either the critical ratio or chi square and demands no assumptions regarding the shapes of the distributions of traits measured by either the criterion or item scores (**26**, pp. 100–102).

In addition to item-discrimination indices based on the significance of differences between percentages in two groups, a number have been proposed for use that are based on the significance of differences between mean criterion scores of two defined groups. One of these is the critical ratio of the difference between mean criterion scores of examinees who do and do not mark a given item choice as correct. When this statistic is computed, Zubin's (**83**) simple correction may be used if the criterion-score variable includes the item for which the critical ratio is being obtained and if the items are scored 1 for "correct" and 0 for "wrong."

Critical ratios may be obtained for differences between the means of defined criterion groups with respect to both the correct response to an item and to the incorrect choices, or decoys. These ratios will not be trustworthy if the number of examinees in each one of the two defined groups that marks it as "correct" is small, say, under 100. This means that calculation of critical ratios for item choices that are not attractive to a fairly large percentage of examinees in even one of the defined groups is questionable unless the total number of examinees is larger than is ordinarily employed.

Cureton has called attention to the advantage of estimating for correct item choices the probability that the mean criterion score of the small group that marks the choice is as large or larger than the mean criterion score of all the examinees tested. The computing procedure he gives would be

$$CR = \frac{M_c - M_t}{s_d},$$

where M_c = mean criterion score of examinees who marked the choice, M_t = mean criterion score of all examinees tested, s_t = standard deviation of criterion scores of all examinees tested, N_t = number of examinees tested, n_c = number of examinees who marked the choice, and

$$s_d = s_t\sqrt{\frac{N_t - n_c}{n_c(N_t - 1)}}.$$

The analogous test for incorrect item choices would be:

$$CR = \frac{M_t - M_c}{s_d}.$$

These tests will be reasonably satisfactory when $N_t \geqq 100$ and

$n_c \geqq 10$ unless the distribution of criterion scores is markedly skewed.

The well-known *t*-test has been used to estimate the probability that examinees who mark a given item choice and those who do not constitute random samples drawn from the same population, but the procedure is laborious.

Recent studies by Walker (**78, 79**) have indicated that items having significant relationships (at stated confidence levels) to a criterion variable can be identified quite rapidly by means of sequential analysis. After the percentage passing an item is obtained, specially prepared tables are used to facilitate the work. The writer was surprised to find how few answer sheets had to be used to establish significance at the 5 per cent level for many items.

A serious limitation of measuring item discrimination in terms of statistical significance lies in the fact that the test constructor is not conveniently able to compare the *amounts* of discriminating power displayed by the items from which he must select those to be included in the final form of a test. Yet, in practice, he is ordinarily faced with the practical problem of selecting the items with the largest amount of discriminating power that will most nearly meet the other requirements for inclusion, such as difficulty level, lack of correlation with some other variable or variables, and subject-matter content. Even if a test constructor were fortunate enough to have a large pool of items available so that he could afford to discard at once all items not meeting some predetermined level of statistical significance, he would be faced with the problem of choosing among the remainder on the basis of amount of discrimination and other requirements, such as those mentioned. In the writer's opinion, therefore, expressing item discriminating power in terms of statistical significance is more satisfying to the statistician than to the test constructor. The latter is better satisfied with appropriate correlation statistics.

Psychophysical methods. The possible application of psychophysical techniques to item analysis was pointed out by Guilford in 1937 (**37**). Methods for accomplishing this were presented in 1942 by Ferguson (**22**) and in 1944 by Finney (**24, 25**), but they have not been widely adopted.

Correlation statistics and their derivatives. A large number of correlation statistics have been proposed for use as measures of item discrimination, and most of them possess certain merits. However, it seems to the writer than unless one makes use of a correlation statistic that is essentially unaffected by variations in item difficulty (or more generally, in the proportions of a given sample marking any item choice) he might

perhaps better use some index of statistical significance. From the writer's point of view, therefore, the biserial and tetrachoric correlation coefficients naturally suggest themselves as the most appropriate statistics for expressing item discrimination. Computation of biserial *r* demands an assumption of normality in the trait measured by the dichotomous scores; computation of tetrachoric *r* demands an assumption of normality in the traits underlying both sets of the dichotomous scores correlated. Therefore, we must recognize that tests of the statistical significance of coefficients computed on the basis of assumptions of normality will be invalidated to the extent that the assumptions are not justified.

So far as the values of the coefficients are concerned, McNemar has pointed out (**58**, pp. 173–174; 178–179) that we can regard them as the values we would expect the product-moment coefficients to take if we had measuring scales for the dichotomized traits that actually did yield normal distributions of continuous variables. Computing formulas for the biserial coefficient and tetrachoric coefficients are given by Davis in a form that incorporates correction for chance success and takes account of adjustments for omitted items and items not reached by some examinees (**18**, formulas 10 and 13). The variance errors of these coefficients can only be approximated, but to test the hypothesis that no correlation exists Davis has presented computing formulas that may be serviceable for those who wish to utilize even a somewhat dubious variance error when it seems desirable to set some sort of objective standard for rejecting items that do not appear to possess significant relationships (at a designated level of confidence).

Unquestionably, biserial correlation coefficients would have been used more frequently in the past for item-analysis purposes if they were not so laborious and, thus, expensive to obtain. To meet the need for an economical approximation to biserial coefficients Flanagan (**30**) has recently prepared special tables. The variance errors of the coefficients read from these tables cannot be obtained analytically, but Flanagan reports that a substantial empirical study has established that the variance errors of biserial *r*'s obtained from the table are very close in magnitude to those computed by formula.

An earlier publication of Flanagan's (**31**) for securing economical approximations to biserial coefficients appeared in 1936.[7] Its use assumes normal continuous distributions of talent underlying the

[7] A new edition of this table with item-difficulty indices incorporated into it in such a way that the item-criterion and item-difficulty indices can be read in one operation is now available through Test Research Service, 12 Normandy Road, Bronxville, N.Y.

dichotomous response categories of "right" and "wrong" as well as rectilinear regression of item scores on criterion scores. It is evident, therefore, that the correlation coefficients estimated by means of Flanagan's table are strictly analogous to tetrachoric correlation coefficients. However, empirical studies made at the Cooperative Test Service of the American Council on Education (**28**) show that their sampling errors are smaller than those of tetrachoric *r*'s and a little larger than those of biserial *r*'s. This may come as a surprise to some research workers who may have supposed that eliminating the middle 46 per cent of the sample would impair the reliability of the resulting data. Additional empirical evidence of the comparative reliability of biserial *r*'s, estimates of the biserial *r*'s derived from Flanagan's table, and tetrachoric *r*'s has been presented by Guilford and Lacey (**38**, pp. 30–31). Internal-consistency item-discrimination indices were computed by several methods for 68 items in a test called Visualization of Maneuvers. Two separate samples of 400 aviation cadets were employed. The reliability coefficient for biserial *r*'s proved to be .87; for estimates read from Flanagan's table, .87; and for tetrachoric *r*'s, .79. Davis, using two comparable samples of 370 aviation students, found the reliability of estimates of biserial *r* read from Flanagan's table to be .67; the standard error of a single biserial coefficient was .08 (**16**, app. A; **21**).

Aschenbrenner has prepared a table analogous to Flanagan's which is entered with percentages of successes in the highest 10 per cent and lowest 10 per cent of the criterion-score distribution (**3**). The use of the table requires exactly the same assumptions as the use of Flanagan's table. Aschenbrenner's table is especially useful when large samples of examinees have been tested and the number of answer sheets to be run through the Graphic Item Counter must be kept rather small, say, approximately two hundred.

When item-discrimination indices are expressed as biserial or tetrachoric correlation coefficients, their interpretation is relatively easy for test constructors adequately trained in statistics. When averages are required or comparisons among items with respect to discriminating ability are to be made, technicians are aware of the precautions that must be observed. Even so, it would often be convenient to have the data expressed on a scale having more nearly the properties of an interval scale. To satisfy the special requirements of test editors for an index of discriminating power that is substantially comparable from item to item and that is easy to use, the writer has suggested indices that constitute a linear function of Fisher's *z* and range from

0 to 99, thus eliminating decimals. They have properties that permit them to be added, subtracted, and averaged more legitimately than other indices; their variance errors are virtually identical regardless of their magnitudes, when they are based on samples of essentially the same size; and the units in which they are expressed are sufficiently coarse to discourage an impression of extreme precision yet fine enough to satisfy all practical requirements of test construction. To minimize the labor required to obtain them, two item-analysis charts have been prepared; one is entered with percentages of successes obtained by examinees in the lowest 27 per cent and the highest 27 per cent of the criterion-score distribution. The other is entered with percentages of successes obtained by examinees in the lowest 10 per cent and the highest 10 per cent of the criterion-score distribution. Each chart yields in one operation for most usable test items both an index of discriminating power and an index of difficulty. The discrimination index has the properties mentioned immediately above and the difficulty indices constitute a linear function of the standard measures corresponding to estimated percentages of successes in the sample tested. The difficulty indices range from 1 to 99 with a median of 50 in each sample of examinees (17).

Three other statistics that have been proposed for use as indices of item discrimination are the correlation ratio (5), the product-moment r, and the phi coefficient (36). If a test constructor is primarily interested in constructing a predictor test for use with a group of examinees having essentially the same level and distribution of ability as the group used for item-analysis purposes, the biserial product-moment r (sometimes called point biserial r) may be used when the criterion is a continuous variable and will be used as such. If the criterion variable is a natural dichotomy and must be used as such, the phi coefficient may be used when the group with which the test is to be used is essentially the same with respect to level and distribution of ability as the group used for item-analysis purposes, and when the point of dichotomy is the same in successive groups in which the test is to be used for prediction purposes.

Computational routines for obtaining product-moment biserial r's and phi coefficients are presented by Kelley (51, pp. 370–373; 377–382). Jurgensen (50) has prepared tables from which phi coefficients may be read directly in the special case when the number of examinees in each dichotomous group is the same. An *abac* published by Guilford (36) makes possible estimation of phi coefficients from percentages of successes in high-scoring and low-scoring groups of a distribution of

criterion scores. Their variance errors are not analytically determinable, however, unless the highest 50 per cent and the lowest 50 per cent of the sample are used.

A convenient method for obtaining item-discrimination indices when the criterion variable is a natural dichotomy has been outlined by Davis (18, pp. 293–295). He recommends the use of computing diagrams prepared by Chesire, Saffir, and Thurstone (14), but diagrams published by Hayes (44) and worksheets designed for use with the latter by Goheen and Kavruck (33) may be employed. Vernon's (76) procedure for obtaining what he calls "double tetrachoric r's" is of interest since these are more reliably determined than ordinary tetrachoric coefficients.

If a dichotomy is enforced upon a continuous series so that the upper and lower 50 per cent or the highest and lowest 25 per cent of the scores are identifiable, computing diagrams prepared by Mosier and McQuitty (61) may be employed to obtain tetrachoric coefficients.

Miscellaneous techniques. Dozens of different kinds of item discrimination indices have been suggested over the past 50 years. Only a few of these have been described in the preceding sections of this paper and space will not permit listing all of them here. Long and Sandiford (56) discussed many of them in 1935, Adkins (1) compared various indices in 1938, Vernon (76) mentioned and evaluated several procedures in 1948, and Gulliksen (41, chap. 21) classified and commented on various techniques in 1950.

Use of Item-Discrimination Data

Improving individual test items. Perhaps the most important use of item-discrimination data is in revising items after tryout. Inspection of choice-by-choice data will ordinarily reveal many incorrect choices that are discriminative in the wrong direction or do not seem to discriminate at all, and some that attract virtually no examinees. These data provide hints regarding the mental processes employed by examinees in answering the items and often lead to insights that enable the test editor to remove ambiguities and to replace invalid or nonfunctioning incorrect choices. This process requires a great deal of ingenuity and perception and is often very time-consuming.

Promoting test homogeneity. Whenever items with high discrimination power are selected in preference to others for the final form of a test, the resulting test tends to be a more homogeneous measure of the first principal component of the items in the tryout test. Unless the latter was carefully constructed to make the first principal component

of the items a satisfactory criterion, selection of the most discriminating items will not necessarily tend to improve test validity.

If homogeneity is a major consideration and the tryout test includes a considerably larger number of items than is needed in the final form, the type of iterative item analysis proposed by Wherry and Gaylord (80) is useful. The procedure calls simply for selecting a group of highly discriminative items and using them to obtain a criterion variable on the basis of which to compute new discrimination indices for all items. The process is then repeated until the items identified as most discriminative change very little with repetition of the process. Methods for selecting homogeneous items have also been developed by Loevinger (54) and Guttman (42). Discussions of these have been prepared by Loevinger (55) and Goodenough (34). Highly homogeneous tests may enhance the efficiency of a battery used for purposes of differential classification, but test validity always remains the primary consideration.

Maximizing test validity. As early as 1934 Horst (47) pointed out that the most effective combination of test items that could be selected would be one in which item intercorrelations would be at a minimum and item correlations with the criterion would be at a maximum. He recognized that the tremendous labor involved in securing data required for the use of classical multiple-regression methods precluded their use. It is now realized that the size of the samples on the basis of which the item intercorrelations and validity coefficients are computed would have to be enormous in order to obtain stable and meaningful regression weights.

Horst (46) described in 1934 a short cut method for approximating the selection of items by multiple-regression procedures. Earlier he had reported the results obtained with this procedure. Rather spectacular validity was secured for a selection test, but the importance of cross-validation was not realized at that time and no validation data on a new sample were secured.

In 1936, Horst (48) proposed another technique for approximating the multiple-regression procedure and Flanagan (29) outlined a sort of iterative procedure for maximizing item validities and minimizing item intercorrelations. Both of these methods made use of the tendency for a reduction in the average internal-consistency discrimination index to be concomitant with a reduction in the average item intercorrelation. The simple process described by Richardson and Adkins (65) in 1938 was also based on this principle.

Toops did a good deal of work on the problem of selecting the most

valid combinations of items and published his methods with Adkins in 1937 and again in 1941, (2, 72). They have not been widely used because of the labor involved. The most recent publication on this matter is that of Gulliksen who presents another approximation technique for determining the combination of items in a tryout test that should yield maximum validity (41, pp. 382–385).

Such experience as the writer has had in trying to maximize the validity of predictor tests, principally the *AAF Qualifying Examination*, has led him to believe that so many factors must be considered in selecting items that strict conformance with procedures developed by Horst and others is rarely practicable. Consequently, he has often simply plotted validity against internal consistency for all of the available items of each type that has shown satisfactory median validity and has then utilized items that satisfy other requirements and that tend to have low internal consistency and high validity for predicting the criterion. This procedure has proved to be practical and would seem to provide safeguards against choosing individual items of high validity but of types that have not displayed satisfactory median validity when a large number of them were validated. Presumably, these safeguards would tend to minimize the shrinkage of test validity coefficients that occurs when tests are validated in similar but different samples of examinees.

Evaluation. An interesting evaluation of the advantages accruing from the use of Horst's maximizing function for selecting items for a predictor test was made several years ago by Gulliksen (40). This indicated some actual gain in validity for a test constructed in accordance with Horst's technique over a test not so constructed. The effectiveness of item selection to maximize test validity depends, of course, on the reliability of the two sets of discrimination indices employed and on the degree of correlation between the total scores on the tryout test and the external variable to be predicted. The higher the reliability of the indices and the lower the correlation of tryout test and external criterion, the greater the advantage that can be gained by this procedure.

It has been the writer's experience that so many factors other than that of item discrimination enter into the selection of items for the final form of a test that discrimination indices do not often play the major role in the process. For achievement tests, great care must be exercised that items judged unacceptable by subject-matter experts be excluded and that the final form preserve the balance among topics specified in the test outline. Then, too, proper regard for the shape of

the distribution of item difficulties must be observed, as noted earlier in this article. The value of item-discrimination indices must always be considered in the light of the adequacy of the criterion variable, the purpose for which the test is to be used, and the way in which it serves that purpose.

The writer knows of no studies that have yielded conclusive evidence that one type of discrimination index is superior to another when each is properly used for selecting items. In fact, it seems likely that the use of different types of indices will lead to the selection of similar items. This does not mean that all sorts of discrimination indices are equally meritorious. Some are apparently less deceptive and more convenient to use than are others. It is obvious that some require far less computational labor than others. The writer's recommendations are largely determined by these considerations.

Summary

Construction of reliable or valid tests that are also efficient measuring instruments requires use of quantitative data concerning the difficulty and discriminating power of each one of the pool of items available.

The measure of item difficulty most commonly used during the past half century has been the one suggested by Binet, namely, the percentage of a defined group passing the item. Refinements designed to permit expressing item difficulty on a more nearly interval scale have been suggested and seem worth while. Until recently, the use of item-difficulty indices to maximize test variance or test validity and to control the standard error of measurement at selected ability levels has not received the attention that this topic warrants.

Many different kinds of item-discrimination indices have been employed. Most of these may be classified as graphic methods, psychophysical methods, methods of expressing the statistical significance of differences, or correlation methods. The writer believes that test constructors will find two correlation statistics (the biserial and tetrachoric coefficients), or approximately linear transformations of them, most useful in the majority of situations.

Procedures for using item-discrimination indices for promoting test homogeneity and for maximizing test validity have been developed. That these two objectives are quite different should be kept constantly in mind. It seems fair to say that there is no conclusive evidence that any one type of item-discrimination data is superior to others for all purposes. It is the writer's belief that the variables used as criteria for

item discrimination are often of doubtful merit; consequently, the usefulness of item-discrimination indices is often smaller than is commonly supposed.

BIBLIOGRAPHY

1. ADKINS, DOROTHY C. A rational comparison of item-selection techniques. *Psychol. Bull.*, 1938, **35**, 655.
2. ADKINS, DOROTHY C., & TOOPS, H. A. Simplified formulas for item selection and construction. *Psychometrika*, 1937, **2**, 165–171.
3. ASCHENBRENNER, R. *A table of values of the product-moment coefficient of correlation as determined by the proportions of successes on the X-variable in each ten-per-cent tail of the Y-variable distribution.* Iowa City: University Examinations Service of the State University of Iowa, 1949.
4. AYRES, L. P. *A measuring scale for ability in spelling.* New York: Russell Sage Foundation, 1915.
5. BARTHELMESS, H. M. The validity of intelligence test elements. *Teach. Coll. Contr. Educ.*, No. 505. New York: Teachers College, Columbia Univ., Bureau of Publications, 1931.
6. BINET, A. La mesure du developpement de l'intelligence chez jeunes enfants. *Bull. Soc. libre pour l'Etude psychol. de l'Enfant.*, 1911.
7. BINET, A. Sur la necessité d'établir un diagnostic scientifique des états inferieurs de l'intelligence. *Année psychol.*, 1905, 11 (Part 1), 163–190.
8. BLISS, E. F. The difficulty of an item. *J. educ. Psychol.*, 1929, **20**, 63–66.
9. BROGDEN, H. E. Variation in test validity with variation in the distribution of item difficulties, number of items, and degree of their intercorrelation. *Psychometrika*, 1946, **11**, 197–214.
10. BROLYER, C. R. The difficulty of a test item. In C. C. Brigham (Chairman), *Third annual report of the chairman on scholastic aptitude tests.* New York: College Entrance Examination Board, 1928.
11. BRYAN, M. M., BURKE, P. J., & STEWART, N. Correction for guessing in the scoring of pretests: effect upon item difficulty and item validity indices. *Educ. psychol. Measmt.* (In press.)
12. CALANDRA, A. Scoring formulas and probability considerations. *Psychometrika*, 1950, **15**, 151–168.
13. CARTER, H. D. How reliable are the common measures of difficulty and validity of objective test items? *J. Psychol.*, 1942, **13**, 31–39.
14. CHESIRE, L., SAFFIR, M., & THURSTONE, L. L. *Computing diagrams for the tetrachoric correlation coefficient.* Chicago: Univ. of Chicago Bookstore, 1933.
15. CLEETON, G. U. Optimum difficulty of group test items. *J. appl. psychol.*, 1926, **10**, 303–326.
16. DAVIS, F. B. *The AAF Qualifying Examination.* AAF Aviation Psychology Program Research Reports, No. 6. Washington, D. C.: Government Printing Office, 1947.
17. DAVIS, F. B. Item-analysis data: their computation, interpretation, and use in test construction. *Harv. Educ. Papers*, No. 2. Cambridge: Graduate School of Education, Harvard University, 1946.
18. DAVIS, F. B. Item selection techniques. In E. F. Lindquist (Ed.), *Educational measurement.* Washington: American Council on Education, 1951.
19. DAVIS, F. B. Notes on test construction: the reliability of item-analysis data. *J. educ. Psychol.*, 1946, **37**, 385–390.

20. DAVIS, F. B. The selection of test items according to difficulty. *Amer. Psychologist*, 1949, **4**, 243. (Abstract)
21. DOPPELT, J. E., & POTTS, E. M. The constancy of item-test correlation coefficients computed from upper and lower groups. *J. educ. Psychol.*, 1944, **40**, 378–381.
22. FERGUSON, G. A. Item selection by the constant process. *Psychometrika*, 1942, **7**, 19–29.
23. FERGUSON, G. A. On the theory of test discrimination. *Psychometrika*, 1949, **14**, 61–68.
24. FINNEY, D. J. The application of probit analysis to the results of mental tests. *Psychometrika*, 1944, **9**, 31–39.
25. FINNEY, D. J. *Probit analysis.* Cambridge: Cambridge Univ. Press, 1947.
26. FISHER, R. A. *Statistical methods for research workers.* (7th Ed.) London: Oliver and Boyd, 1938.
27. FLANAGAN, J. C. General considerations in the selection of test items and a short method of estimating the product-moment coefficient from the data at the tails of the distributions. *J. educ. Psychol.*, 1939, **30**, 674–680.
28. FLANAGAN, J. C. Item analysis by test scoring machine graphic item counter. In *Proc. educ. Res. Forum.* Endicott: International Business Machines Corp., 1940. Pp. 89–94.
29. FLANAGAN, J. C. A short method for selecting the best combination of test items for a particular purpose. *Psychol. Bull.*, 1936, **33**, 603–604.
30. FLANAGAN, J. C. *A table for obtaining the biserial correlation coefficient.* Pittsburgh: American Institute for Research, 1950.
31. FLANAGAN, J. C. *A table of the values of the product-moment coefficient of correlation in a normal bivariate population corresponding to given proportions of successes.* New York: Cooperative Test Service, 1936.
32. GIBBONS, C. C. The predictive value of the most valid items of an examination. *J. educ. Psychol.*, 1940, **31**, 616–621.
33. GOHEEN, H. W., & KAVRUCK, S. A worksheet for tetrachoric *r* and standard error of tetrachoric *r* using Hayes' diagrams and tables. *Psychometrika*, 1948, **13**, 279–280.
34. GOODENOUGH, W. H. A technique for scale analysis. *Educ. psychol. Measmt.*, 1944, **4**, 179–190.
35. GUILFORD, J. P. The determination of item difficulty when chance success is a factor. *Psychometrika*, 1936, **1**, 259–264.
36. GUILFORD, J. P. The phi coefficient and chi square as indices of item validity. *Psychometrika*, 1941, **6**, 11–19.
37. GUILFORD, J. P. The psychophysics of mental test difficulty. *Psychometrika*, 1937, **2**, 121–133.
38. GUILFORD, J. P., & Lacey, J. I. (Eds.) *Printed classification tests.* AAF Aviation Psychology Program Research Reports, No. 5. Washington, D. C. Government Printing Office, 1947.
39. GULLIKSEN, H. O. The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, 1945, **10**, 79–92.
40. GULLIKSEN, H. O. *Selection of test items by correlation with an external criterion, as applied to the Mechanical Comprehension Test, OQT, 0-2.* OSRD Publications Board, No. 13319. Washington, D. C.: Department of Commerce, 1946.
41. GULLIKSEN, H. O. *Theory of mental tests.* New York: Wiley, 1950.
42. GUTTMAN, L. A basis for scaling qualitative data. *Amer. sociol. Rev.*, 1944, **9**, 139–150.
43. HAMILTON, C. H. Bias and error in multiple-choice tests. *Psychometrika*, 1950, **15**, 151–168.

44. HAYES, S. P., JR. Diagrams for computing tetrachoric correlation coefficients from percentage differences. *Psychometrika*, 1946, **11**, 163–172.
45. HORST, A. P. The difficulty of a multiple-choice test item. *J. educ. Psychol.*, 1933, **24**, 229–232.
46. HORST, A. P. Increasing the efficiency of selection tests. *Personnel J.*, 1934, **12**, 254–259.
47. HORST, A. P. Item analysis by the method of successive residuals. *J. exp. Educ.*, 1934, **2**, 254–263.
48. HORST, A. P. Item selection by means of a maximizing function. *Psychometrika*, 1936, **1**, 229–244.
49. JACKSON, R. W. B., & Ferguson, G. A. A plea for a functional approach to test construction. *Educ. psychol. Msmt.*, 1943, **3**, 23–28.
50. JURGENSEN, C. E. Table for determining phi coefficients. *Psychometrika*, 1947, **12**, 17–29.
51. KELLEY, T. L. *Fundamentals of statistics*. Cambridge: Harvard Univ. Press, 1947.
52. KELLEY, T. L. The selection of upper and lower groups for the validation of test items. *J. educ. Psychol.*, 1939, **30**, 17–24.
53. LAWLEY, D. N. On problems connected with item selection and test construction. *Proc. roy. Soc. Edin.*, 1942–1943, **61** (Section A, Part III), 273–287.
54. LOEVINGER, JANE. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, **61**, (4), Whole No. 285.
55. LOEVINGER, JANE. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychol. Bull.*, 1948, **45**, 507–529.
56. LONG, T. A., & SANDIFORD, P., *et al.* *The validation of test items*. Bulletin No. 3. Toronto: Department of Educational Research, Ontario College of Education, 1935.
57. LYERLY, S. B. A note on correcting for chance success in objective tests. *Psychometrika*, 1951, **16**, 21–30.
58. MCNEMAR, Q. *Psychological statistics*. New York: Wiley, 1949.
59. MOLLENKOPF, W. G. An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, 1950, **15**, 291–315.
60. MOLLENKOPF, W. G. Variation in the standard error of measurement. *Psychometrika*, 1949, **14**, 189–229.
61. MOSIER, C. I., & MCQUITTY, J. V. Methods of item validation and abacs for item-test correlation and critical ratio of upper-lower difference. *Psychometrika*, 1940, **5**, 57–65.
62. MURPHY, H. D., & Davis, F. B. Note on the measurement of progress in remedial reading. *Peabody J. Educ.*, 1949, **27**, 108–111.
63. PATERSON, D. G. *Preparation and use of new-type examinations*. Yonkers: World Book Co., 1926.
64. RICHARDSON, M. W. The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1936, **1**, 33–49.
65. RICHARDSON, M. W., & ADKINS, DOROTHY C. A rapid method of selecting test items. *J. educ. Psychol.*, 1938, **29**, 547–552.
66. RUCH, G. M. *The objective or new-type examination*. New York: Scott, Foresman, 1929.
67. SYMONDS, P. M. Factors influencing test reliability. *J. educ. Psychol.*, 1928, **19**, 73–87.
68. THORNDIKE, E. L. *The measurement of intelligence*. New York: Teachers College, Columbia Univ., Bureau of Publications, 1927.
69. THURSTONE, L. L. Scale construction with weighted observations. *J. educ. Psychol.*, 1928, **19**, 441–453.

70. THURSTONE, L. L. The unit of measurement in educational scales. *J. educ. Psychol.*, 1927, **28**, 505–524.
71. THURSTONE, THELMA G. The difficulty of a test and its diagnostic value. *J. educ. Psychol.*, 1932, **23**, 335–343.
72. TOOPS, H. A. The L-method. *Psychometrika*, 1941, **6**, 249–266.
73. TUCKER, L. R. Maximum validity of a test with equivalent items. *Psychometrika*, 1946, **11**, 1–13.
74. TURNBULL, W. W. A normalized graphic method of item analysis. *J. educ. Psychol.*, 1946, **37**, 129–141.
75. UHRBROCK, R. S., & RICHARDSON, M. W. Item analysis. *Personnel J.*, 1933, **12**, 141–154.
76. VERNON, P. E. Indices of item consistency and validity. *Brit. J. Psychol.* (Stat. Sect.), 1948, **1**, 152–166.
77. VOTAW, D. F. Notes on validation of test items by comparison of widely spaced groups. *J. educ. Psychol.*, 1934, **25**, 185–191.
78. WALKER, HELEN M. Item selection by sequential sampling. *Teach. Coll. Rec.*, 1949, **50**, 404–409.
79. WALKER, HELEN M., & COHEN, S. *Probability tables for item analysis by means of sequential sampling.* New York: Teachers College, Columbia Univ., Bureau of Publications, 1949.
80. WHERRY, R. J., & GAYLORD, R. H. Factor pattern of test items and tests as a function of the correlation coefficient, content, difficulty and constant error factors. *Psychometrika*, 1944, **9**, 237–244.
81. WISSLER, C. *The correlation of mental and physical tests.* New York: Columbia Univ. Press, 1901.
82. WOODY, C. Measurements of some achievements in arithmetic. *Teach. Coll. Contr. Educ.*, No. 80, New York: Teachers College, Columbia Univ., Bureau of Publications, 1916.
83. ZUBIN, J. The method of internal consistency for selecting test items. *J. educ. Psychol.*, 1934, **25**, 345–356.

Received July 23, 1951.

NON-PARAMETRIC STATISTICS FOR PSYCHOLOGICAL RESEARCH

LINCOLN E. MOSES
Teachers College, Columbia University

It has been said that "everybody believes in the law of errors [the normal distribution], the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."[1] There are excellent theoretical reasons to explain the preeminent position which the normal distribution has held in the development of statistical theory.[2]

On the other hand, at some time or other nearly every experimenter must work with samples which he knows do not come from a normal distribution. If he knows what the distribution actually is then he may find a transformation such that his transformed data *are* observations from a normal distribution, or he may find a special theory already worked out (as for, say, the Poisson distribution). More often he has no such knowledge of the population distribution, and then he must choose between applying the textbook methods in violation of their underlying assumptions, or of finding valid techniques which have no underlying assumptions concerning the shape of the parent population.

Until about fifteen years ago this was merely Hobson's choice, since about the only distribution-free methods were rank correlation and χ^2 tests. But there has recently been a great growth in statistical methodology which provides the experimenter with tools free of assumptions about the population distribution. These techniques are generally referred to as Non-Parametric Methods, or sometimes, Distribution Free Methods.

It is the purpose of this paper to present some of the principal methods, and an intuitive explanation of their rationale, properties, and applicability, with a view to facilitating their use by workers in psychological research.

In many, but not all, of the methods discussed, the data to which the tests are applied are not the original measurements in the sample

[1] Cramér cites Poincare's quotation of this famous remark by Lippman. (CRAMÉR, H., *Mathematical methods of statistics*, Princeton Univ. Press, 1946, p. 232.)

[2] Perhaps the most outstanding of these is the so-called Central Limit Theorem which specifies (roughly) that means of "large" (enough) samples from *any* population (except for some pathological cases which cannot occur in practice) are normally distributed. The discussion in this paper is not concerned with the "large sample case."

but simply their ranks, or perhaps only their signs. This feature of the methods arouses some criticism. It is intuitively obvious that a statistical procedure that replaces each of the two sets of numbers below by the *same* set of plus and minus signs:

$-8, -3, -2.1, -1.6, .3$	$- - - - +$
$-14, -14, -9.4, -.2, 5.7$	$- - - - +$

is not using all the information the data provide—is "throwing away" information. This is a less telling indictment than it seems to be. The relevant question is not "How much information does a certain statistical procedure throw away?", but rather "Of the methods available—classical, or non-parametric—which best uses the information in the sample?" Since the answer to the question will depend on the sort of population from which the sample comes, no general answer can be given. In the literature of mathematical statistics (**3**, **23**, **26**) there are examples of distributions where a non-parametric test which "throws away information" is clearly superior to a *t*-test, for instance. How the comparison would work in any given case is a matter of conjecture. The following advantages and disadvantages of the non-parametric methods should be considered:

Advantages of non-parametric methods:

1. Whatever may be the form of the distribution from which the sample has been drawn, a non-parametric test of a specified significance level actually *has* that significance level (provided that the sample has been drawn at random; in certain cases as will be noted, it is also necessary to assume that the distribution is continuous).

2. If samples are very small, e.g., six, there is in effect no alternative to a non-parametric test (unless the parent distribution really is known).

3. If the sample consists of observations from several *different* populations there may be a suitable non-parametric treatment.

4. The methods are usually easier to apply than the classical techniques.

5. If the data are inherently of the nature of ranks, not measurements, they can be treated directly by non-parametric methods without precariously assuming some special form for the underlying distribution.

6. In certain cases data can only be taken as "better" or "worse," that is, an observation can only be characterized as a plus or minus. Obviously, the classical tests are not directly applicable to such data.

Disadvantages of non-parametric methods:

1. If non-parametric tests rather than normal-theory tests are applied to normal data then they are wasteful of data. The degree of wastefulness is measured by the "efficiency" of the non-parametric test. If, for example, a test has 80 per cent efficiency this means that *where the data are from a normal distribution*, the appropriate classical test would be just as effective with a sample of 20 per cent smaller size. The efficiency thus expresses the relative

merits of the non-parametric test and the classical test under the conditions *where the normal test is correct*, but does not tell us how the tests will compare on non-normal data.

2. The non-parametric tests and tables of significance values are widely scattered in the periodical literature.

3. For large samples some of the non-parametric methods require a great amount of labor, unless approximations are employed.

Tests Based on Plus or Minus

The Sign Test. One of the best known and most widely applicable of the techniques to be discussed in this paper is the statistical sign test. A complete treatment will be found in (2). In many cases where an experimenter wishes to establish that two treatments are different (or that a particular one of the two is better) he is able to employ matched pairs, one member of each pair being assigned (at random) to treatment A, the other to treatment B. The classical technique is to apply a t-test to the differences; the underlying assumptions are that the differences are normally distributed with the same variance. The assumptions underlying the sign test are simply: (a) that the variable under consideration has a continuous distribution and (b) that both members of any pair are treated similarly—except for the experimental variable. There is an assumption neither of normality nor of similar treatment of the various pairs. Thus the different pairs may be of different socio-economic status, age, IQ, etc., so long as within each pair such relevant extranea are comparable. The hypothesis tested is that "The median difference is zero."[3] The test is performed by considering the differences $X_{Ai}-X_{Bi}$ and noting whether the sign is plus or minus. If the null hypothesis is true we expect about an equal number of plus and minus signs. The hypothesis is rejected if there are too few of one sign. The probability level of any result can be evaluated by the binomial expansion with $p=\frac{1}{2}$ and $N=$the number of pairs. Tables of significance values for various sample sizes are available (1). A table of sample sizes necessary to detect with probability .95 a departure from the null hypothesis of various degrees (e.g., that $P(X_A>X_B)=.3$) at significance levels .01, .05, .10, .25 is given by Dixon and Mood (2). For $N\geqq 30$ the normal approximation to the binomial will suffice.

If it is desired to test not merely that treatments A and B differ,

[3] The hypothesis is also properly expressed:

$$P(X_A > X_B) = P(X_B > X_A) = \tfrac{1}{2}$$

This is read in words: the probability that X_A will exceed X_B is equal to the probability that X_B will exceed X_A (and thus equal to $\frac{1}{2}$). X_A and X_B are two members of a pair.

but that treatment A is actually better than treatment B, a significant result can arise only if the number of minus signs is too small.

An extension of the sign test will permit one to determine whether A is better than B by, say, 5 points. Formally the null hypothesis is: $P(X_A > X_B + 5) \leqq \frac{1}{2}$.[4] In this case one considers the differences $X_{Ai} - (X_{Bi} + 5)$ and rejects the null hypothesis for too few minus signs.

Another extension enables one to determine whether A is better than B by some specified percentage—say 10 per cent. The null hypothesis is: $P(X_A > 1.10X_B) \leqq \frac{1}{2}$. If a significantly small number of the differences $X_{Ai} - (X_{Bi})$ (1.10) are negative the null hypothesis is rejected. Both these extensions are applicable only where the numbers are additive and the second is legitimate only if there is a zero point on the scale.

The efficiency of the sign test (in the sense defined) declines from around 95 per cent for $N = 6$ (25) to 62 per cent for very large samples. Where data are easily gotten, the extraordinary simplicity of computation sometimes justifies taking a larger sample and using the sign test, even though the classical methods would be justified and more efficient. In certain cases there is no substitute for the sign test, as where a pair of protocols can be assessed as to which exhibits more "cooperation" but there is little hope of a numerical evaluation.

The Median Test. In some cases where two treatments (or groups) are to be compared as to whether they are drawn from populations having the same median (or to determine whether a particular one of the two populations has a smaller median), it is not possible to work with matched pairs. The hypothesis can be tested by the median test (16, p. 394). The samples need not be of equal size. Suppose there are n X's and m Y's. Compute the median for the combined sample of $n+m$ observations. If the samples do come from populations with the same median then we should expect about half of the X's to be above the common median and about half below, similarly for the Y's. If the relative proportions are too discrepant, we reject the hypothesis of equality.

To perform the test, record a plus for any observation above the common median, a minus for any observation below the median. Then construct a 2×2 contingency table. For instance, suppose that an experiment yielded the following data:

Control Group: (X) 10, 15, 13, 12, 12, 14, 11, 9
Experimental Group: (Y) 7, 7, 8, 6, 13, 9

[4] In words this is read: the probability is at most $\frac{1}{2}$ that X_A will exceed X_B by 5. The null hypothesis can also be expressed: the median difference $X_A - X_B$ is equal to at most 5 in the population.

All observations greater than or equal to 11 are +'s; all 10 or less are −'s.

	+	−
Experimental Group	1	5
Control Group	6	2

The significance of the data is evaluated in the same manner as if this were a 2×2 test of independence. For such small frequencies as these Fisher's exact method must be used (5, Sec. 21.02); for large enough frequencies χ^2 with one degree of freedom is the test statistic, Yates's correction being used unless the number of cases is large. If the hypothesis is being tested against an alternative on one side only, i.e., the question asked of the data is not "are the two medians equal," but "is $\mathrm{Md}(X) \geqq \mathrm{Md}(Y)$," the ordinary techniques associated with χ^2 and one-sided test apply.

The assumptions underlying this test are that the X's and Y's are random samples from their respective populations, and that the population distributions are of the same form, differing only by a translation up or down the scale. Although the test is derived using the second assumption, Mood states that the test "is sensitive primarily to differences in location and very little to differences in shape."

Tests Based on Rank Order

There is a group of important methods which deal with the data in terms of their ranks. Four of the most important will be discussed here: a rank test for matched pairs (27, 28); the "T" test of Wilcoxon for two unmatched samples (27, 28), together with its extension by Mann and Whitney (13); the analysis of variance by ranks (6); the run test.

Wilcoxon's Matched Pairs Signed Ranks Test. Where the experimenter has paired scores X_{Ai} under treatment A and X_{Bi} under treatment B, he can rank the differences *in order of absolute size;* he may be unable to give numerical scores to the observations in each pair and still be able to rank the differences in order of absolute size. The ranking is done by giving rank 1 to the numerically *least* difference, rank 2 to the next least, etc. If methods A and B are equivalent, that is, if there is no difference and the null hypothesis is true, he should expect some of the larger, and some of the smaller, absolute deviations to arise with A being superior, some with B superior. That is, the sum of the ranks where A is favored should be about equal to the sum of the ranks where

B is favored. If the sum of the ranks for the negative differences is too small, or if the sum of the ranks for the positive differences is too small, the null hypothesis is to be rejected. Tables of significance values for the smaller sum of ranks will be found in (27) for n (the number of pairs) equal to 7 through 16. Tables for n from 7 to 25 are available in (29). For $n \geqq 25$ the sum of ranks T may be taken as normally distributed with mean $= \overline{T} = n(n+1)/4$ and standard deviation $\sqrt{(2n+1)\overline{T}/6}$. For example: suppose that seven pairs of rats are divided into a control and an experimental group. Suppose that the data are their times to run a certain maze and are as shown in Table 1.

TABLE 1

ILLUSTRATIVE DATA FOR TEST OF SIGNIFICANCE USING WILCOXON'S MATCHED PAIRS SIGNED RANKS TEST

Pair	*Exp.*	*Control*	*Diff. Exp-Control*	*Rank Diff.*	*Ranks with Less Frequent Sign*
(a)	65	51	14	6	
(b)	60	44	16	7	
(c)	71	64	7	4	
(d)	52	55	−3	1	1
(e)	62	49	13	5	
(f)	43	38	5	2	
(g)	58	52	6	3	
Total of ranks with less frequent sign					1

First, it is worth noting that these data are amenable to treatment by the sign test. Six of the differences have the same sign. The probability of six or more signs alike, if in fact the median difference is zero, is equal to $16/128 = 1/8$. Therefore, these data would not be regarded as cause for rejection using the sign test. But a closer examination of the data shows not only that there was only one negative difference but that it was the smallest difference in the set. These data argue more strongly against the null hypothesis than would the same set of differences with, say, pair (*e*) being the sole negative difference (or indeed any other one difference) though any of these possible samples would be treated identically by the sign test. It turns out that application of the rank test under consideration will adjudge these data as significant; essentially the different answer arises from exactly the considerations just sketched—the *size* of the sole negative difference is taken into account.

Wilcoxon's tables tell us that for $n=7$ a rank total of 2 or less for one of the groups is significant at level .05, and the null hypothesis of equality of treatments is rejected. The tables referred to are for two-sided tests. If one desires to test a one-sided hypothesis he may use the .05 level to determine a test of significance level .025, provided that the observed values lie in the direction of rejecting the one-sided hypothesis. Similar remarks apply to other significance levels.

A confidence interval for the difference in the treatment effects can be obtained as follows. Suppose that to the time of every rat in the control group $4\frac{1}{2}$ seconds were added, then all the differences would have the same sign as at present, the ranks would be the same, and the treatments would still be adjudged as significantly different. However, if $5\frac{1}{2}$ seconds were added to each control group score the groups would not differ significantly. The boundary for this argument occurs at 5. Similarly, if $14\frac{1}{2}$ be added to all the control group times the differences become all negative except for pair (*b*) which is then $+1\frac{1}{2}$ having a rank of 2.5 (it is tied with (*e*) for second and third place); this gives a "smaller rank total" of 2.5 which is not significant. But if 14 2/3 were added to each control group score then (*b*) would be the lone positive difference with rank 2; this would be significant. Since alterations in the differences greater than 5 and less or equal to 14.5 do not yield a significant difference, but values outside this range do, we can take 5 to $14\frac{1}{2}$ as a 95 per cent confidence interval for the increase in running time associated with the experimental treatment.

Mann-Whitney "U" Test. Where the observations are not made on matched pairs, but two unmatched groups are to be compared, the Mann-Whitney "*U*" test (or in the case of equal sized groups, its equivalent, the Wilcoxon "*T*" test) for two samples can be applied.

The null hypothesis which is tested is that the two groups of observations—say n X's and m Y's—have been drawn from a common population (that is, "there is no difference"). The test is designed to detect (roughly stated) whether one population has a larger mean than the other. Precisely stated, it is designed to guard against the alternative hypotheses that for every a, $P(X>a) \geqq P(Y>a)$ or $P(X>a) \leqq P(Y>a)$. A special case (unnecessarily restrictive) is where X and Y are assumed to have the same distribution except for a translation along the scale, so that the X's are all smaller—or all larger—than the "corresponding" Y's; here the null hypothesis says that there is no translation at all, and the test has the property that if in fact there is a positive or negative translation, then with a sufficiently large sample the test will reject the null hypothesis with any desired degree of probability.

To apply the test one arranges the $m+n$ observations in increasing order of size (algebraic sign not being ignored) and substitutes their ranks (1 for the smallest, $m+n$ for the largest). If the two samples were of equal size, so that $m=n$, the sum of the ranks for the X's should about equal the sum of the ranks for the Y's under the null hypothesis. If $m \neq n$ then the sums would be roughly proportional to the sizes m and n. The test consists in determining whether the observed discrepancy is too large to have arisen reasonably by chance, with the null hypothesis being true.

Tables of significance values for all possible pairs of sample sizes with $m \leqq 8$, $n \leqq 8$ are given in (13). For m and n both greater than 8 the test statistic is nearly normally distributed and the test of significance is made by employing this fact. If $m, n \geqq 8$, then U is normally distributed with mean $mn/2$ and standard deviation $\sqrt{mn(n+m+1)/12}$; one has merely to rank the $m+n$ observations from least to greatest, find T, the sum of the Y ranks, and from this calculate U, and see whether it is too many standard deviations removed from its expected value, $mn/2$.

TABLE 2

ILLUSTRATION FOR THE MANN-WHITNEY "U" TEST

Variable	*Observation*	*Rank*
X	10.2	1
X	12.8	2
Y	13.4	3
X	13.5	4
X	16.0	5
Y	17.1	6
Y	17.3	7
X	18.0	8
X	18.2	9
X	19.0	10
Y	19.4	11
X	19.5	12
Y	21.3	13
Y	24.0	14

$\sum X$ ranks $= 51$
$T = \sum Y$ ranks $= 54$.

As an example, suppose that there were 8 X's and 6 Y's, so that $m=6$, $n=8$ and that the data arranged in order of size were as shown in Table 2. The tables of significance are given in terms of U where

$$U = mn + \frac{m(m+1)}{2} - T.$$

Here $U = 6 \times 8 + (6 \times 7/2) - 54 = 15$.

The table for $n = 8$ tells us that a U as small as 15 has a probability level of .141; so that the null hypothesis is accepted.

In using these tables the reader will find that the probability of small values of U is given. To find the probability $U \geqq k$ where k is a number larger than those given in the tables he uses the identity:

$$P(U \geqq k \mid n X\text{'s}, m\ Y\text{'s}) = P(U \leqq mn - k \mid n X\text{'s}, m\ Y\text{'s}).$$

It is further to be noted that m and n are entirely symmetrical, so that $P\{U = k \mid n\ X\text{'s}, m\ Y\text{'s}\} = P\{U = k \mid m\ X\text{'s}, n\ Y\text{'s}\}$.

As an example, suppose that an experimenter has 5 X's and 8 Y's and that the sum of the Y ranks, T, is 39. Then

$$U = 6 \times 8 + \frac{8(8+1)}{2} - 39 = 45.$$

This is a large value of U and is not tabled; to decide whether or not it is significantly large we note that

$$P(U \geqq 45) = P(U \leqq 6 \times 8 - 45) = P(U \leqq 3).$$

The tables tell us that this probability is .002.

Analysis of Variance with Ranked Data. The assumptions underlying the analysis of variance are: the observations are independent; they are drawn from normal populations all of which have the same variance; the means in these normal populations are linear combinations of "effects" due to row and/or columns, etc., that is, effects are additive.

Correlation among the observations would be perhaps the most dangerous assumption failure; but careful design should usually eliminate this. In some cases both normality of distribution and homogeneity of variance can be approximated either in the data, or by some transformation. In other cases this cannot be done. The analysis of variance by ranks is a very easy procedure and does not depend on such assumptions. It has the further advantage of enabling data which are inherently only ranks to be examined for significance.

Let there be n replications of an experiment where each subject undergoes a different one of p treatments. In each replication there are

a different p subjects. Data from such an experiment might be as follows:

TABLE 3

ILLUSTRATION FOR ANALYSIS OF VARIANCE WITH RANKED DATA

	Treatment				
	A	*B*	*C*	*D*	*E*
Group 1	11(2)	14(4)	13(3)	9(1)	20(5)
Group 2	12(3)	11(2)	13(4)	10(1)	18(5)
Group 3	16(3)	17(4)	14(2)	13(1)	19(5)
Group 4	9(1)	11(3)	14(4)	10(2)	16(5)
Rank totals	9	13	13	5	20

The numbers appearing in parentheses are the ranks from least to greatest *within each row* (replication). If the treatments A, B, C, D, E ($p=5$) are not different, then the rank totals would be expected to turn out about equal. In the present example there seems to be a marked disparity. To evaluate its significance we compute the statistic χ_r^2, done below, which has approximately the χ^2 distribution with $p-1$ degrees of freedom.

$$\chi_r^2 = \frac{12}{np(p+1)} \times \text{Sum (rank totals)}^2 - 3n(p+1)$$

Here $n=4$, $p=5$ and the statistic becomes:

$$\chi_r^2 = \frac{12}{120}\cdot(844) - 12(6)$$

$$= 12.4$$

For 4 degrees of freedom this is significant at level .02 but not .01.

If the groups 1, 2, 3, 4 in the example themselves represented four treatments, or age levels, etc., then a test of the equality of those four treatments could also be made by interchanging rows and columns. For that test χ_r^2 would have 3 degrees of freedom since then $p=4$, $n=5$.

A full treatment of the mathematical basis for the test is given by Friedman (6). Kendall and Smith (12) give exact probabilities for small m and n, and a detailed consideration of the closeness of approximation and recommendations for evaluation of significance levels are given in

Friedman's article (7). Wilcoxon (29) gives several instructive illustrations showing, among other things, how interactions can be tested.

Wald-Wolfowitz Run Test. The final test employing ranked data which will be considered is the Wald-Wolfowitz run test. This is a test of the hypothesis that two samples (not necessarily of equal size) have been drawn from a common population. It has the property that if the X's and Y's are not from a common population then, no matter in what way the populations differ (dispersion, median, skewness, etc.) the test will—for sufficiently large samples—reject the null hypothesis with probability as near to 1 as is desired. The application of the test is extremely simple.

Just as for the U test, arrange the combined sample of m Y's and n X's in increasing order. Then a run is defined as a sequence of letters of the same kind which cannot be extended by incorporating an adjacent observation. Thus there are 9 runs below:

$$\underline{X_1\ X_2}\ Y_1\ \underline{X_3}\ Y_2\ Y_3\ Y_4\ Y_5\ \underline{X_4\ X_5}\ Y_6\ \underline{X_6\ X_7\ X_8}\ Y_7\ \underline{X_9\ X_{10}}$$

The X runs are underlined; the Y runs stand between them.

Now if the two samples are from a common population then the X's and Y's will generally be well mixed and the number of runs will be large. But if the X population has a much higher median, then there is to be expected a long run of Y's at one end, a long run of X's at the other, and consequently a reduced total number of runs. If the X's come from a population with much greater dispersion then there should be a long run of X's at each end, and a reduced total number of runs. Similar arguments apply to opposite skewness, etc. Generally, then, rejection of the null hypothesis will be indicated if the runs are too few in number. An important application of the run test is to test randomness of grouping; in some such cases ***either*** too many or too few runs might be basis for rejection. A nice example is given by Swed and Eisenhart (24) where the question at issue is, are seats at a lunch counter a half hour before the rush hour occupied at random? Very many runs of occupied and empty seats would clearly be an a priori cause for rejection. So would too few runs if the possibility of friends coming together was to be considered. In the example to which the U test was earlier applied, only too few runs would be reasonable cause for rejection if the X's and Y's represented, say, examination scores for two different statistics classes.

The run test can also be applied to a series of events ordered in time. Let there be n observations arranged in order of the time at which they were taken. Let those greater than the median be denoted by X, those

less than the median by Y. If one suspects a time trend—like gradual increase—or a "bunching" in time due to change in attitude, etc., he would reject for too few runs.

Tables of significance for the run test are given by Swed and Eisenhart (**24**), for $m, n \leqq 20$. For larger samples the number of runs d can be taken as being normally distributed with mean $= (2mn/m+n)+1$ and

$$\text{standard deviation} = \sqrt{\frac{2mn(2mn - m - n)}{(m + n)^2(m + n - 1)}}.$$

Mood (**16**) states that for practical purposes this approximation will suffice for $m, n \geqq 10$. To apply this large sample theory one merely decides before taking the sample whether rejection is indicated by too many, or too few, (or either) runs and then sees whether d is too many standard deviations removed from its expected value in the rejection direction.

Mathematical investigations of this test indicate that because it guards against *all* kinds of difference between the distribution functions of X and Y it is not very powerful against any particular class of alternatives. Thus, if one were interested in detecting whether one population had a greater median than another he would do better to employ a test such as the U test. A related point is that when one rejects the null hypothesis on the basis of the run test, he can assert that the two populations differ—but he has little if any clue as to *how* they differ. Often the purpose is to establish that there is a difference in *means*, or *dispersion*, and the run test gives an answer which is not easy to interpret.

The only assumption involved in the run test is that the common population is continuous. This assumption is involved in all the tests depending on rank presented here. Generally, if there is a small number of ties the average rank for each set of tied observations may be given to each and the test carried through.

Randomization Tests

There is a variety of non-parametric tests which employ the numerical values of the data themselves. Among the most important of these are techniques based on the method of randomization. This kind of test was proposed by Fisher (**4**, Sec. 21), and has received extended treatment and development by Pitman (**21**, **22**).

Matched Pairs. All the randomization tests are based on parallel logic. The simplest with which to exhibit the rationale is the matched

pair case. Suppose, for example, that we have two observations (one under condition A, the other under condition B) on each of seven individuals. The null hypothesis is that conditions A and B are no different; the data are shown in Table 4. The average difference is

TABLE 4

ILLUSTRATION FOR RANDOMIZATION TEST USING MATCHED PAIRS

i	X_{Ai}	X_{Bi}	$d_i = X_{Ai} - X_{Bi}$
1	14.9	15.5	−.6
2	17.3	16.5	.8
3	14.9	13.2	1.7
4	18.1	16.0	2.1
5	12.0	12.1	−.1
6	19.4	18.1	1.3
7	15.6	11.4	4.2
			9.4 = S

1.34, but is it significantly different from zero at, say, the 5 per cent level? To answer this with the t-test we would assume that the differences were normally distributed with a common unknown variance. We can get an exact test assuming only that the d_i are random samples from a common population. If the null hypothesis is true, then conditions A and B are experimentally indistinguishable, and for any individual the distinction between his X_A and X_B is merely a convention of labelling; in particular, the difference $X_{A3} - X_{B3} = 1.7$, say, is just exactly as likely as that $X_{B3} - X_{A3} = 1.7$. This means that associated with this sample are many other possible ones, all of which (under the null hypothesis) were exactly as likely to occur as this. For instance, the sample might just as well have turned out: +.6, −.8, −1.7, −2.1, +.1, −1.3, +4.2 or +.6, +.8, +1.7, +2.1, +.1, −1.3, −4.2, etc. In all, there are $2^7 = 128$ such outcomes, all equally likely under the null hypothesis that the treatments A and B are experimentally indistinguishable. With each of these is associated an $S = \Sigma d_i$. Some of these 128 S's are just about what one would expect if the null hypothesis were true, i.e., near zero. A few are well removed from zero—and much like what we expect under an alternative hypothesis such as the population mean of A exceeding that of B—or vice versa; we write these $\mu_A > \mu_B$ and $\mu_B > \mu_A$ in the sequel. To get an exact test of, say, level .05, we select of the samples which we can thus generate, that 5 per cent of them most likely under the alternatives we wish to guard against,

and constitute these chosen possible samples as our rejection region. In the present case, $.05(128) = 6.4$, so we choose six possibilities. The probability of getting one of these six samples under the null hypothesis is $6/128 = .047$. Then if the sample we actually drew is one of these listed for the rejection region we reject the hypothesis of equality of A and B. In our numerical example, if the investigator's "experimental hypothesis" had been: condition B leads to larger scores on the average than does condition A, he would test the null hypothesis of equality of A and B but would reject it only if the d_i were predominantly negative. If they were predominantly positive *or* well balanced he would have to regard the data as failing to support his experimental hypothesis. His rejection region would be six samples giving the greatest negative S. If he actually desired only to determine whether the two conditions yield different average scores then he must regard either a large positive S or a large negative S as cause for rejection, and his rejection region would consist of the three samples yielding greatest $+S$ and the negatives of these samples, which will yield the greatest $-S$.

Let us find the two-sided region just described. If all the d_i were positive then S would be 10.8, its maximum value. The next largest possible value for S (10.6) would be where all but $d_5 = .1$ were positive. Such considerations lead to the following list of the first 5 positive samples in order of size of S:

							S
.6	.8	1.7	2.1	.1	1.3	4.2	10.8
.6	.8	1.7	2.1	−.1	1.3	4.2	10.6
−.6	.8	1.7	2.1	.1	1.3	4.2	9.6
*−.6	.8	1.7	2.1	−.1	1.3	4.2	9.4
.6	−.8	1.7	2.1	.1	1.3	4.2	9.2

Thus the sample we obtained, which has been starred, lies in the acceptance region and the null hypothesis stands. If, however, only the alternative $\mu_A > \mu_B$ was being tested against, then the top six positive values would be the 5 per cent rejection region, and our sample, being 4th, would lie in it.

For large n, say 20, the number of possible samples which we can generate by altering signs on the given numbers is large ($2^{20} > 1{,}000{,}000$) and even listing a 1 per cent rejection region is a massive undertaking. There are two principal alternatives. Wilcoxon's T test, where ranks are substituted for numbers, may be used (in fact, the T test may be regarded as a randomization test on the ranks—and this clue should enable the reader to find the one-sided significance points for the T test where n is small). The second alternative is that where $n \geqq 12$

(roughly), and where the d_i are of roughly the same size (as a rule it might be safe to require $(d_k^2/\Sigma d_i^2) \leqq 5/2n$, where d_k is the largest difference in the set) a normal approximation can be used.

Each d_i, under the null hypothesis, is a chance variable taking the values $\pm d_i$ each with probability $\frac{1}{2}$. The d_i are independent. One form of the central limit theorem ensures that under the conditions given, the exact distibution of $z = S/\sqrt{\sum d_i^2}$ in the "randomization distribution" will be very closely approximated by the unit normal distribution. This test is obviously easy to apply. On data which are in fact normally distributed it is 100 per cent efficient for large samples. Examples of non-normal populations can be given where despite this efficiency it is an inferior test as compared with the rank T test (which has large sample efficiency of 95 per cent).

Two Sample Test. A randomization significance test for two samples has the same underlying logic. Let there be n X's and m Y's. If there is "no difference" then the fact that in the pooled ordered sample *a particular* n observations are labelled X is, so to speak, one of many equally likely accidents. All together there are

$$\binom{m+n}{n} = \frac{(m+n)!}{m!\ n!}$$

equally likely ways in which the relabelling might be done. For certain of these the "spread" or difference between ΣX and ΣY is extreme. The construction of the test consists in choosing a number k of these for a rejection region. If α is the significance level then k is chosen so that

$$k = \alpha\binom{m+n}{n},$$

as nearly as is possible. The choice of *which* k most extreme possible outcomes should constitute the rejection region depends, as always, on what alternatives are to be guarded against.

An example follows:

X	Y
11.6, 12.1, 12.2, 12.6, 13.1	9.5, 10.7, 11.8

We test $\mu_X = \mu_Y$ against the alternative $\mu_X > \mu_Y$.

The arithmetic is made more convenient if from all numbers we subtract 9.5, and then multiply by 10. We now have:

21 26 27 31 36	0 12 23

The average of these eight numbers is 22. If, then, the null hypothesis is true we should expect to find ΣY near (3) (22) = 66. In all there are

$$\binom{5+3}{3} = \frac{8.7.6}{1.2.3} = 56$$

possible equally likely samples. If we are working at level .05 we shall choose the 2 samples (out of the 56) most likely under the alternative hypothesis $\mu_X > \mu_Y$. These are, obviously

23 26 27 31 36 0 12 21

and

21 26 27 31 36 0 12 23

The second of these is the sample we obtained, and the null hypothesis is rejected. For illustrative purposes the six most extreme (two-sided) samples are listed below:

X	Y	$\sum Y - 66$
23 26 27 31 36	0, 12, 21	33 − 66 = 33
0 12 21 23 26	27, 31, 36	94 − 66 = 28
21 26 27 31 36	0, 12, 23	35 − 66 = 31
0 12 21 23 27	26, 31, 36	93 − 66 = 27
21 23 27 31 36	0, 12, 26	38 − 66 = 28
0 12 21 26 27	23, 31, 36	90 − 66 = 24

If m and n are large, the carrying out of these computations becomes essentially impossible. But again there exists a convenient approximation to the distribution of the statistic in the randomization distribution of

$$\binom{m+n}{n}$$

possible sample values.

Provided that:

(1) $1/4 \leqq (m/n) \leqq 4$.

(2) $(\mu_4/\mu_2^2) - 3$ (The kurtosis computed for the pooled sample), not large; then the following statistic has approximately the t distribution with $m+n-2$ degrees of freedom:

$$\frac{\overline{Y} - \overline{X}}{\sqrt{\dfrac{\sum (Y - \overline{Y})^2 + \sum (X - \overline{X})^2}{m+n-2}\left(\dfrac{1}{m} + \dfrac{1}{n}\right)}}$$

It is a curious result; this is the ordinary t statistic. This means that provided conditions 1 and 2 hold (and these can be checked from the

sample) the t statistic actually gives a test of the stated significance level *without* the usual assumptions being a part of the inference. It has *not* been assumed that X and Y are normally distributed with a common variance.

If the distributions of X and Y both have finite variance, but different means, the probability that the test will reject the null hypothesis tends to one as both m and n become large. If the distributions are different but have the same mean this is not so.

An alternative to the use of the t statistic to approximate the randomization distribution is to employ the Mann-Whitney U test. There are circumstances under which the U test (though it "throws away" data by reducing the observations to ranks) is the better test. The U test may be regarded as test of the randomization type applied to the *ranks* of the observations.

Confidence Intervals. In both these cases (paired or unpaired observations) confidence intervals can be obtained by adding equal increments to one set of values until a significant positive difference is first reached, and then altering them still further until a significant negative difference is first reached. These two extreme alterations constitute the end points of a confidence interval for the true difference. If the approximations (normal, and t) are to be used, then the conditions for their validity must hold at these extreme points; otherwise the exact procedure has to be used.

Correlation and Tests of Independence. The problem of correlation can also be attacked by the randomization method. That is, one can test the hypothesis of zero correlation with samples of small (or large) size without making assumptions about the form of the joint distribution of X and Y. For a treatment of the problem, see (**22**).

Tests of Independence

When one has a pair of observations (X_i, Y_i) for each member of his sample and desires to test the independence of X and Y there are numerous techniques available. The rank-order correlation coefficient, or τ, Kendall's rank-order statistic (**11**), may be used. The product-moment coefficient can be tested non-parametrically as mentioned in the preceding paragraph.

In addition there is an extraordinarily easily applied method, Olmstead and Tukey's corner test of association (**20**). Its efficiency and other properties await a full mathematical investigation, but informed opinion holds that it is likely to be a very good test.

To apply the test one first plots the observations in a scatter diagram. Then, following simple rules given by Olmstead and Tukey (**20**) and also by Mood (16), the statistician measures the degree to which the data are concentrated in the "corners" of the scatter diagram. (The instructions referred to essentially define the "corners.") If a substantial number of observations are concentrated in diagonally opposite corners (which would be expected in the presence of strong association between the variables), then the null hypothesis of independence is rejected. Although the use of this technique is simple, the explanation of how to construct the corners is rather lengthy and will be omitted here. The test is entirely distribution-free. Because of its ease of application it should find frequent use as a preliminary test to determine whether a product-moment correlation coefficient is worth computing in cases where the latter is fully justified.

There also exist non-parametric methods for linear regression, including tests of significance. They will not be taken up here, but a full treatment of both their mathematical theory and method of application is given by Mood (16, ch. 16). In this source there is also a technique for analysing one and two factor experiments; an alternative to the analysis of variance by ranks. All these methods depend upon the way in which the medians of various subclasses behave. They are all completely distribution free. As an attack on the analysis of variance problem they are more flexible than analysis of variance by ranks, but are less efficient, and probably not to be preferred for problems of an uncomplicated design.

Percentiles

If one has a sample of n observations and wishes to estimate the percentiles of the parent distibution he will, of course, employ the percentiles in the sample. Confidence intervals (confidence coefficient $1-\alpha$) may be obtained as follows. If the sample is arranged in order of increasing size:

$$X_1, X_2, X_3, \cdots, X_n$$

then X_1 is the smallest observation, X_2 the next smallest, etc. Let ξ_p denote the 100 p percentile. Then

$$P(X_r < \xi_p < X_s) = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}$$

Using tables of the binomial distribution (19), one then chooses r and s so that the probability (the value of the sum) is at least $1-\alpha$.

If there are ties in the data then

$$P(X_r < \xi_p < X_s) \geqq \sum_{i=r}^{s-1} \binom{n}{i} p^i - (1-p)^{n-i}$$

For example, a .90 confidence for the 40th percentile in the population from which this sample comes:

17	21	23	24	24	35	27	30
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8

is:

$$17 \quad \text{to} \quad 27,$$

that is:

$$X_1 \quad \text{to} \quad X_7.$$

Since:

$$\sum_{i=1}^{6} \binom{8}{i} (.4)^i (.6)^{8-i} = .90,$$

where we have $p = .4$, $n = 8$.

For large samples the binomial sum can be approximated by the normal distribution. The index i is approximately normal (for large n) with mean np and standard deviation $\sqrt{npq}$. So to obtain a 95 per cent confidence interval one would count 1.96 $\sqrt{npq}$ observations to the right and to the left of the $100p$ sample percentile to find the observations whose numerical values constitute a 95 per cent confidence interval.

Some Other Non-Parametric Methods

Certain important topics in the field of non-parametric methods have been either completely omitted, or merely mentioned in this paper. Among these are:

Rank Correlation Methods. A recent book by Kendall (11) provides the experimenter with a rather generous variety of techniques not elsewhere published. Among other matters of interest considered there are: tied ranks, coefficient of concordance (with significance test) to measure agreement among more than two judges, significance of the difference between two non-zero rank-order correlation coefficients. Work is being done in this field by Kendall and his associates and additional results will be published in the near future.

Kolmogorov-Smirnov Tests. These tests serve as alternatives (preferable for certain reasons) to χ^2 for two classes of problems:

1. To test the hypothesis that a random sample has been drawn from a population with a certain specified distribution.
2. To test that two random samples (of not necessarily equal size) have been drawn from the same population. The methods apply only where the chance variable is continuous. An excellent non-mathematical discussion, with tables and examples, is given by Massey (14). Some more recent results and tables for the two-sample problem are also given by Massey (15).

Tests for Randomness of a Sequence of Numerical Observations. The Wald and Wolfowitz run test discussed in this paper is one test of this sort, where two groups of observations are involved. Where there is only one sequence of observations, perhaps ordered in time, one may still wish to know whether they may be regarded as a random sequence. An informative non-mathematical discussion of this problem, with several tests, is found in Moore and Wallis (17).

Tolerance Intervals. One can ask the question: "Between what limits can I be nearly sure (say 95 per cent, or 99 per cent, etc.) that at least 90 per cent (or 80 per cent, or 98 per cent, etc.) of the population values lie?" These limits are called tolerance limits. The problem clearly differs from the confidence interval problem, which is concerned with location of the population mean, or a certain population percentile, etc.

A brief discussion will be found in Dixon and Massey (1). Some useful charts which eliminate computations are given by Murphy (18), where the relevant literature is also cited.

Literature on Non-Parametric Methods

The textbook literature presents few extended treatments of non-parametric methods. Of those known to the writer, one of the fullest, and surely the least mathematical, is Chapter 17 of Dixon and Massey's text (1). For the reader with facility in advanced calculus many important methods are explained and derived in Chapter 16 of Mood's text (16). At a mathematical level intermediate between these two is Chapter 8 of Johnson's text (9) and Chapter 9 of Hoel's text (8). Finally, the mathematically mature reader will find many of the techniques taken up in this paper (and some others) discussed in somewhat greater detail in Chapter 21, Volume II of Kendall's advanced book (10).

A paper by S. S. Wilks (30) affords a complete but terse review of the whole field up through about 1947. The treatment requires a good knowledge of mathematical statistics. A full bibliography is included.

The following bibliography is not intended to be complete. The reader who wishes to explore any one topic in detail will find little difficulty in uncovering the relevant literature with the aid of the references cited in the papers listed here.

BIBLIOGRAPHY

1. Dixon, W. J., & Massey, F. J., Jr. *Introduction to statistical analysis.* New York: McGraw-Hill, 1951.
2. Dixon, W. J., & Mood, A. M. The Statistical Sign Test. *J. Amer. statist. Ass.*, 1946, **41**, 557–566.
3. Festinger, L. The significance of difference between means without reference to the frequency distribution function. *Psychometrika*, 1946, **11**, 97–106.
4. Fisher, R. A. *Design of experiments.* London: Oliver and Boyd, 1936.
5. Fisher, R. A. *Statistical methods for research workers.* London: Oliver and Boyd, 1925.
6. Friedman, Milton. Use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. statist. Ass.*, 1937, **32**, 675–701.
7. Friedman, Milton. A comparison of alternative tests of significance for the problem of *m* rankings. *Ann. math. Statist.*, 1940, *11*, 86–92.
8. Hoel, P. G. *Introduction to mathematical statistics.* New York: Wiley, 1947.
9. Johnson, P. O. *Statistical methods in research.* New York: Prentice-Hall, 1949.
10. Kendall, M. G. *The advanced theory of statistics*, Vol. II. London: C. Griffin and Co., 1948.
11. Kendall, M. G. *Rank correlation methods.* London: C. Griffin and Co., 1948.
12. Kendall, M. G., & Smith, B. B. The problem of *m* rankings. *Ann. math. Statist.*, 1939, **10**, 275–287.
13. Mann, H. B., & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. math. Statist.*, 1947, **18**, 50–60.
14. Massey, F. J., Jr. The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. statist. Ass.*, 1951, **46**, 68–78.
15. Massey, F. J., Jr. The distribution of the maximum deviation between two sample cumulative step functions. *Ann. math. Statist.*, 1951, **22**, 125–128.
16. Mood, A. M. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1950.
17. Moore, G. H., & Wallis, W. A. Time series significance tests based on signs of differences. *J. Amer. statist. Ass.*, 1943, **38**, 153–164.
18. Murphy, R. B. Non-parametric tolerance limits. *Ann. math. Statist.*, 1948, **19**, 581–589.
19. National Bureau of Standards. *Tables of the binomial probability distribution.* Washington, D. C.: U. S. Government Printing Office, 1949.
20. Olmstead, P. S., & Tukey, J. W. A corner test for association. *Ann. math. Statist.*, 1947, **18**, 495–513.
21. Pitman, E. J. G. Significance tests which may be applied to samples from any population. *Suppl. J. Royal statist. Soc.*, 1937, **4**, 119.
22. Pitman, E. J. G. Significance tests which may be applied to samples from any population, II. The correlation coefficient test. *Suppl. J. Roy. statist. Soc.*, 1937, **4**, 225.
23. Pitman, E. J. G. Notes on non-parametric statistical inference. (Unpublished.)
24. Swed, F. S., & Eisenhart, C. Tables for testing randomness of grouping

in a sequence of alternatives. *Ann. math. Statist.*, 1943, **14**, 66–87.

25. WALSH, J. E. On the power of the sign test for slippage of means. *Ann. math. Statist.*, 1946, **17**, 358–362.
26. WHITNEY, D. R., *A Comparison of the power of non-parametric tests and tests based on the normal distribution under nonnormal alternatives.* Unpublished Ph.D. dissertation at Ohio State University, 1948.
27. WILCOXON, FRANK. Individual comparisons by ranking methods. *Biometrics Bull.*, 1945, **1**, 80–82.
28. WILCOXON, FRANK. Probability tables for individual comparison by ranking methods. *Biometrics*, 1947, **3**, 119–22.
29. WILCOXON, FRANK. *Some rapid approximate statistical procedures.* American Cyanamide Co., 1949.
30. WILKS, S. S. Order statistics. *Bull. Amer. math. Soc.*, 1948, **54**, 6–50.

Received July 19, 1951.

A COMPUTATIONAL SHORT CUT IN FACTOR ANALYSIS

R. V. ANDREE
University of Oklahoma

One of the more onerous computations in the use of factor analysis, linear programming, and other analyses is the computation of inverse matrices. Recently a method was published which appreciably shortens this chore.[1] The portion of this paper of interest to psychologists is presented here.

Let us first prove the following:

THEOREM: If A is a non-singular square matrix, then for any P and Q such that $PAQ=I$, $A^{-1}=QP$.

If A, P, and Q are non-singular matrices such that

$$PAQ = I,$$

then

$$QPAQ = QI = Q.$$

By multiplying on the right by Q^{-1}, one obtains

$$QPA = I,$$

or

$$QP = A^{-1}.$$

A simple method of obtaining a P and Q, employed in the theorem, is developed. It is well known that if A^{-1} exists, then A can be carried into the identity matrix I by elementary transformations on rows and columns. The author constructs an L-shaped array by placing identity matrices of the same size as A to the left and below A as follows:

$$L = \begin{matrix} I & A \\ & I \end{matrix}\Bigg].$$

If one performs the elementary transformations (adding k times one row to another row, etc.) in this array, carrying A into I, a suitable P and Q are obtained. It must be kept in mind that all transformations must occur *within A;* the other matrices are merely used to keep cumulative track of the result of these transformations.

$$L = \begin{matrix} I & A \\ & I \end{matrix}\Bigg] \rightarrow \begin{matrix} P & I \\ & Q \end{matrix}\Bigg].$$

In any computational process it is essential that a method exist for

[1] ANDREE, R. V. Computation of the inverse of a matrix. *Amer. math. Month.*, 1951, 58, 87–92.

checking the process at frequent stages. The above process can be checked at any stage desired. Assume we have obtained

$$L = \begin{matrix} I & A \\ I & \end{matrix} \Bigg] \rightarrow \begin{matrix} N & A' \\ & M \end{matrix} \Bigg].$$

To check the work one need only verify that $NAM = A'$.

An example may clarify the method. Let us compute A^{-1} where

$$A = \begin{bmatrix} 17 & 44 & 26 \\ -5 & -28 & 9 \\ 8 & 19 & 14 \end{bmatrix}.$$

First form

$$L = \left[\begin{array}{ccc:ccc} 1 & 0 & 0 & 17 & 44 & 26 \\ 0 & 1 & 0 & -5 & -28 & 9 \\ 0 & 0 & 1 & 8 & 19 & 14 \\ \hdashline & & & 1 & 0 & 0 \\ & & & 0 & 1 & 0 \\ & & & 0 & 0 & 1 \end{array}\right].$$

By row and column transformations on A, reduce A to I. By working in the L-shaped array we keep track of the result of the transformations made, thus determining a P and Q for our theorem.

Rather than divide the first row (or fourth column) of the array by 17, I have chosen to subtract twice the third row from the first to obtain a 1 in the a_{11} position of A.

Thus

$$L \rightarrow \left[\begin{array}{ccc:ccc} 1 & 0 & -2 & 1 & 6 & -2 \\ 0 & 1 & 0 & -5 & -28 & 9 \\ 0 & 0 & 1 & 8 & 19 & 14 \\ \hdashline & & & 1 & 0 & 0 \\ & & & 0 & 1 & 0 \\ & & & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc:ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 5 & 1 & -10 & 0 & 2 & -1 \\ -8 & 0 & 17 & 0 & -29 & 30 \\ \hdashline & & & 1 & -6 & 2 \\ & & & 0 & 1 & 0 \\ & & & 0 & 0 & 1 \end{array}\right].$$

Adding the last column to the next to last, and continuing, one obtains

$$\left[\begin{array}{ccc:ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 5 & 1 & -10 & 0 & 1 & -1 \\ -8 & 0 & 17 & 0 & 1 & 30 \\ \hdashline & & & 1 & -4 & 2 \\ & & & 0 & 1 & 0 \\ & & & 0 & 1 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc:ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 5 & 1 & -10 & 0 & 1 & 0 \\ -13 & -1 & 27 & 0 & 0 & 31 \\ \hdashline & & & 1 & -4 & -2 \\ & & & 0 & 1 & 1 \\ & & & 0 & 1 & 2 \end{array}\right].$$

Upon dividing the third row by 31, one obtains

$$\left[\begin{array}{ccc:ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ 5 & 1 & -10 & 0 & 1 & 0 \\ -\frac{13}{31} & -\frac{1}{31} & \frac{27}{31} & 0 & 0 & 1 \\ \hdashline & & & 1 & -4 & -2 \\ & & & 0 & 1 & 1 \\ & & & 0 & 1 & 2 \end{array}\right] = \left[\begin{array}{cc} P & I \\ & Q \end{array}\right].$$

By the use of our theorem,

$$A^{-1} = QP = \begin{bmatrix} 1 & -4 & -2 \\ 0 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & -2 \\ 5 & 1 & -10 \\ -\frac{13}{31} & -\frac{1}{31} & \frac{27}{31} \end{bmatrix} = \begin{bmatrix} -\frac{563}{31} & -\frac{122}{31} & \frac{1124}{31} \\ \frac{142}{31} & \frac{30}{31} & -\frac{283}{31} \\ \frac{129}{31} & \frac{29}{31} & -\frac{256}{31} \end{bmatrix}.$$

By using a different sequence of transformations one obtains a different P and Q, but the product QP is again A^{-1}; for example,

$$\left[\begin{array}{ccc:ccc} 1 & 0 & -2 & 1 & 0 & 0 \\ \frac{5}{2} & \frac{1}{2} & -5 & 0 & 1 & 0 \\ \frac{129}{31} & \frac{29}{31} & -\frac{256}{31} & 0 & 0 & 1 \\ \hdashline L\rightarrow & & & & & \\ & & & 1 & -6 & -1 \\ & & & 0 & 1 & \frac{1}{2} \\ & & & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc} P & I \\ & Q \end{array}\right],$$

and

$$QP = \begin{bmatrix} 1 & -6 & -1 \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -2 \\ \frac{5}{2} & \frac{1}{2} & -5 \\ \frac{129}{31} & \frac{29}{31} & -\frac{256}{31} \end{bmatrix} = \begin{bmatrix} -\frac{563}{31} & -\frac{122}{31} & \frac{1124}{31} \\ \frac{142}{31} & \frac{30}{31} & -\frac{283}{31} \\ \frac{129}{31} & \frac{29}{31} & -\frac{256}{31} \end{bmatrix} = A^{-1}.$$

The author also obtains a step process for computing A^{-1} by machine. However, since the computer who understands the method is better able to take advantage of fortunate combinations which may occur, the step process will be omitted here.

One of the well-known "short cut methods" for computing matric inverses is a special case of the method presented here, in which A is transformed into I by row transformations only. In actual practice it is often shorter to use both row and column transformations, and then form the product QP as suggested in this paper. This is particularly true when some of the elements are integers (whole numbers) or when the elements of a given column or row are pre-known factors of one another or differ by a pre-known amount, as is often the case with actual laboratory data.

Received July 26, 1951.

NOTE ON "A QUALIFICATION IN THE USE OF ANALYSIS OF VARIANCE"

C. H. PATTERSON[1]

Veterans Administration Center, Fort Snelling, St. Paul, Minnesota

In a recent article (5) Webb and Lemmon raise the problem of the occurrence of discrepancies between the results of the *F*-test in the analysis of variance and the application of the *t*-test for individual comparisons after the *F*-test has been applied to the data.

A complete answer to the problems posed by Webb and Lemmon would require more space than the original article, and the writer does not deem himself competent to attempt this. However, certain rather obvious points should be made to correct the erroneous impression given by the article in question. In order to limit the present remarks to the proportions of a note, I shall confine myself to the enumeration of certain considerations which would appear to be pertinent.

1. With the exception of Fisher, the writers list only secondary sources who are not accepted as authorities in statistics. They were apparently not aware, for example, of Tukey's (4) article, or of other references quoted by him. Moreover, if they had read further in Fisher, they would have found, following the paragraph which they quote, a suggested solution to one of the problems which they raise (see section 4 below).

2. The writers correctly state the assumptions of the analysis of variance, i.e., equivalence of variance and randomness of selection, but actually the examples which they give are not random occurrences, nor even selected actual occurrences, but hypothetical, fictitious data.

3. A consideration of the principles of randomness and probability would explain some of the apparent contradictions found by Webb and Lemmon. For example, in their first illustration, Case I, with two groups, F and t are in agreement, since for one degree of freedom $t^2 = F$. Then, for Case II, a third group is added—not randomly—and the authors appear to be surprised that F and t do not agree. This third group is selected to be midway between the other two, or at the grand mean, thus not contributing to the between-group sum of squares. F is not significant, and the authors suggest that since the first two groups have not changed, t is significant. Actually, of course, the first two groups are no longer comparable without consideration of the third group.

That these results are not in disagreement, as would appear, is shown

[1] I wish to express my appreciation to Dr. P. O. Johnson of the University of Minnesota, who read a draft of the manuscript and made a number of helpful suggestions.

by a consideration of the nature of the problem. If the selection of groups were actually random, the probability of selecting two such extreme groups out of two choices would be *less* than the probability of selecting two such extreme groups out of three choices, which is Case II. Therefore, it would be expected that a difference in significance would exist, since the probabilities are different. Having added the third group, the changed probabilities must be considered in any comparisons. It is precisely the value of the *F*-test and the analysis of variance that these differing probabilities are considered, and all the differences evaluated or tested simultaneously.

4. Actually, the results of the *F*-test and the *t*-test are not in disagreement if the *t*-test is applied properly. Cochran and Cox (**2**) recognize the problem posed by Webb and Lemmon, pointing out that "in order that the *F*- and *t*-tests be valid, the tests to be made in an experiment should be chosen before the results have been inspected" (p. 67), and indicating that, when this condition is not met, by the ordinary criterion of significance "the effect is to obtain too many significant results, or to raise the significance level of the test from the presumed 5% to some higher level, usually unknown" (p. 68).

Fisher (**1**, pp. 57–58) has suggested the appropriate modification of the *t*-test, and it is illustrated in Johnson (**3**, p. 234). Essentially, it consists in adjusting the level of significance required, to take into account the changed probabilities involved. In Case II of Webb and Lemmon, for example, the required probability (at the 1 per cent level) would not be 1 in 100, but 1 in 3 (the number of possible comparisons) $\times 100$, or 1 in 300. An alternative procedure which has been suggested would be to use a standard error of the difference based on all three samples, multiplying by $\sqrt{2}$ to adjust for the fact that the comparison is not of random, but of selected means. In both cases, the *t*-test and the *F*-test would be in agreement. These tests are intuitive, but they do take into consideration the unique probabilities involved. Tukey (**4**) suggests other methods for isolating significantly different subgroups or single means when the *F*-test indicates that the means are not alike.

The number of comparisons which can be made in any normal population is very large, and many differences would be found to be significant by the standards of probability associated with the assumption of independent and random samples. The fundamental consideration is whether or not the design of experiment was set up to test a particular comparison, or whether it was discovered to be of interest after the results were obtained. In the latter case, as both Fisher and Cochran and Cox indicate, the level of significance must be adjusted to take into account the different probabilities.

5. Cases III, IV, and V of Webb and Lemmon introduce additional factors. Actually, these are problems of covariance, but they are not considered as such. Furthermore, even if they were so treated, the

assumption of linearity of regression is violated. These illustrations are therefore not pertinent to the problems which they raise, and, as they suggest, "it would probably be more appropriate to test for the significance of regression or trend" (5, p. 135) by methods applicable to nonlinear functions.

REFERENCES

1. FISHER, R. A. *The design of experiments.* 5th ed. London: Oliver & Boyd, 1949.
2. COCHRAN, W. G., & COX, G. M. *Experimental designs.* New York: John Wiley, 1950.
3. JOHNSON, P. O. *Statistical methods in research.* New York: Prentice-Hall, 1949.
4. TUKEY, J. W. Comparing individual means in the analysis of variance. *Biometrics,* 1949, **5,** 99–114.
5. WEBB, W. B., & LEMMON, V. W. A qualification in the use of analysis of variance. *Psychol. Bull.,* 1950, **47,** 130–136.

Received July 26, 1951.

COMMENT ON "A QUALIFICATION IN THE USE OF ANALYSIS OF VARIANCE"

SOLOMON DIAMOND
Los Angeles State College of Applied Arts and Sciences

Webb and Lemmon have contended, in this JOURNAL[1], that analysis of variance does not offer an appropriate test for a true difference among the means of several groups, and have offered an hypothetical example in which the existence of such a true difference is presumably masked by inclusion of data on a third group, in an over-all analysis of variance. This comment is intended to point out two fallacies in their argument. An empirical demonstration will also be given, by a sampling experiment with random numbers.

Starting with two groups, A and B, whose means on the Y-variable "are just significantly different at the 5 per cent level as shown by the t-test," they add a third group, C, selected to occupy a position midway between A and B; on the X-variable. Assuming linear relationship between X and Y, they assert that "the mean of C would lie on the grand mean of all three groups," that including C would therefore make no contribution to the between-groups sum of squares despite adding one degree of freedom, and that under these conditions the F-ratio would fail to reach the critical level. Thus, adding C has masked the difference between A and B; by implication, the analysis of variance does not offer an appropriate test in such situations.

The first fallacy in this argument is the assumption that the mean of C will fall on the regression line established by A and B. On the contrary, if no inappropriate restrictions are operating in the gathering of data, the mean of C surely will not fall on that line. This is a simple consequence of the fact that experimental data are always fallible, and the means of samples cannot be expected to coincide with the true means of the populations from which they are drawn. (If this were not so, we would have no need for analysis of variance or any other sampling theory.) Therefore, the mean of C will not coincide with the grand mean, and, what is equally important, the grand mean will no longer lie midway between the means of A and B. In consequence, not only does C make some contribution of its own to the between-groups sum of squares, but its inclusion causes a necessary increase in the sum of the joint contributions of A and B. These additions tend to raise F, and

[1] WEBB, W. B., & LEMMON, V. W. A qualification in the use of analysis of variance. *Psychol. Bull.*, 1950, 47, 130–136.

constitute an essential factor in the situation, which was overlooked by Webb and Lemmon.

It is possible, however, that what Webb and Lemmon expect may occur: although *A* and *B* alone show a significant difference, the *F* based on these together with *C* may be too low to permit rejection of the null hypothesis. If this happens, shall we hold by the indication of the *t*-test? Before deciding this question, we must consider the second fallacy.

This is the implicit assumption that when a "just significant" *t* was obtained, an error of Type II was avoided (accepting the null hypothesis when in fact it was false). When including *C* leads to a contrary conclusion, Webb and Lemmon feel that this Type II error is being thrust upon them. However, no sampling theory preserves us from all error. It is possible that the original rejection of the null hypothesis represented an error of Type I (rejecting the null hypothesis when it was in fact true). Since *t* was "just significant," there was a 5 per cent chance that such an error occurred. Without additional data, there is no way of determining whether we have escaped an error or been led into one. As additional data become available, we must permit them to sway our judgment. If including *C* results in an *F* below the critical value, we must accept the implication that we had probably committed an error of Type I, which has been discovered. A "just significant" result is like a balance on a knife-edge. A change in one score in either group will tip the balance in the direction of greater or lesser significance. This risk that additional data may reverse our judgment is inherent in any result of borderline significance.

A sampling experiment. The thesis that including intermediate groups makes it more difficult to recognize a significant difference can be readily tested in a sampling experiment. The writer has performed such an experiment, conforming as closely as possible to the conditions of Webb and Lemmon's hypothetical illustration. They stipulate that each group has an *N* of 11, that they have equal variance, and that *t* for the difference between the means of *A* and *B* is the critical value, 2.086. If groups of this size have a mean difference of 2, and yield an estimated population variance of 5, the resulting *t* would be 2.098, which is an adequate approximation. Our experiment was performed by drawing random samples, each with an *N* of 11, from a population with a variance of 5. The population consisted of the 10,000 four-place numbers, which were first distributed in a normal distribution with variance 5. (The interval of 1 therefore represents somewhat less than half of a standard deviation.) Fifteen successive samples, each consisting of 11 four-place numbers, were drawn from a table of random numbers. These 15 samples were then randomly assigned to the classes *A*,

B, and *C*. Sums of scores and sums of squares were computed for each sample, as if the central score of the parent distribution had been 5 for the *A*'s, 7 for the *B*'s, and 6 for the *C*'s.

Since we have five *A*'s and five *B*'s, we can make 25 *t*-tests for the significance of the difference between the *A* and *B* populations. Since an ideally representative *A* and an ideally representative *B* would yield a just significant *t*, we should expect about half of our pairs to indicate a significant difference. Actually, only 8 of the 25 do so. So great a departure from the expected mean value would occur about once in ten times by chance.

The crucial part of our sampling experiment requires that we include one of the *C*-groups as a third member in each set, and then apply analysis of variance. Each of the five *C*-groups was used in combination with each *A*-*B* pair, making 125 such tests in all. The results were as follows:

1. In no case did the addition of an intermediate group produce a significant *F*, where the *A*-*B* pair involved did not yield a significant *t*.
2. In both cases in which the *t*'s had been significant at the 1 per cent level, all *F*'s were likewise significant at the 1 per cent level.
3. In four of the six cases where *t*'s had reached the 5 per cent but not the 1 per cent level, all *F*'s were at the 5 per cent level; in one case, one *F* (out of five) missed the 5 per cent level; *in one case, no significant F's were obtained.*

What would be the effect of including all five *C*-groups with each of the *A*-*B* pairs, making a series of tests in each of which seven classes are involved? By an extension of the reasoning used by Webb and Lemmon, the masking effect should be even more pronounced. (They do extend the argument in this way, in their Case III, which we shall not consider in detail.) Actually, in these tests, we found that where the *A*-*B* pair included had yielded a *t* significant at the 1 per cent level, the *F* based on seven classes was likewise significant at the 1 per cent level; but *F*'s significant at the 5 per cent level were found for only four of the six pairs that had yielded *t*'s at that level.

TABLE 1

NUMBER OF SIGNIFICANT TESTS FOUND IN EACH SERIES

Groups included	$p < .01$	$.01 < p < .05$	$p < .05$
A-*B* pairs	2	6	17
C_1 and *A*-*B* pairs	2	5	18
C_2 and *A*-*B* pairs	2	4	19
C_3 and *A*-*B* pairs	2	5	18
C_4 and *A*-*B* pairs	2	5	18
C_5 and *A*-*B* pairs	2	5	18
All *C*'s and *A*-*B* pairs	2	4	19

The results, which are summarized in Table 1, show that including intermediate groups does not have the drastic masking effect that Webb and Lemmon anticipated. Nevertheless, there remains the *one pair which yielded a significant t, but no significant F*'s when included with any of the *C* groups, or with all of them together, in analysis of variance. Does this one case justify their position? Let us examine it more closely. The mean difference found between *A* and *B* in this pair of samples is not 2.0, which is the difference between the population means, but 1.64. The estimates of population variance based on these samples are low: 2.7 and 3.7. An investigator who had independent knowledge of the fact that these samples were drawn from populations with a variance of 5, would be very reluctant to accept a conclusion based on these two samples. If the same investigator lacked such independent knowledge, but was confronted instead by the fact that including one or more intermediate groups in an analysis of variance led to a nonsignificant *F*, he would be rash to suppose that the test based on two groups was more sound than the one based on three or more. Thus, this case is the exception that proves the rule: including *C* reversed our judgment, but did so on wholly valid grounds.

We conclude that there is no reason to doubt the appropriateness of analysis of variance as a test for a significant difference among means of several groups. There is no reason to suppose that such a test is more liable to error than the *t*-test of two groups, provided that the basic assumption of homogeneity of variance is satisfied. To make a separate test for the significance of the difference between the two extreme groups in such an experimental design is not permissible, unless justified by some very special condition. One such condition might be a demonstration that homogeneity of variance does apply to the populations from which the extreme groups are drawn, but does not apply to the intermediate groups—but this is only another way of saying that we can question the conclusion from the analysis of variance only if the basic conditions for a valid analysis of variance have not been met.

Received July 31, 1951.

A SEQUEL TO THE NOTES OF PATTERSON AND DIAMOND

WILSE B. WEBB AND VERNON LEMMON
Washington University

After reading the articles by Diamond and Patterson, we find ourselves in the position of agreeing with a good many of their statements, while at the same time feeling that most of their criticisms have not been aimed directly at the core of our problem. Most of their discussion seems to deal with problems of the conventional analysis of variance situation, in which the means of the groups show no trend, but are randomly related to each other. Under these circumstances, we stoutly agree that an estimate of significance based on a few groups should be revised when more groups are added, and we agree that it is not legitimate, when the *F*-test fails to show significance, to single out two groups to compare by means of the *t*-test after the experiment has been completed. We further agree that sin is bad.

We were concerned, however, with a fairly common experimental design in which the means of the groups show a definite ordering or trend, with respect to some other experimental variable. To quote, "In each case the location of the means was not a matter of chance, but was due to a functional relationship with some other variable." In this special case, we wondered whether the conventional analysis of variance technique would give the best information on significance. Perhaps, if the trend is real, even a small difference between two means should be called significant, in spite of the findings of the conventional tests. Our critics do not seem to have devoted much attention to this problem.

One comment which seems pertinent is Patterson's suggestion that if the regression is linear, covariance analysis might be employed. This is consistent with the last statement in our article, just before the summary, to the effect that in the case of ordered variables, a test of regression or trend might come closer to yielding the kind of information needed.

The attitude of the present authors is simply summarized: In each and every application of a statistic, the interpretation of this statistic is inherently dependent upon a logical analysis of the experimental situation involved. Adherence to this dictum does not seem too clearly a part of the comments of Diamond or Patterson.

Received September 12, 1951.

HANDBOOK OF EXPERIMENTAL PSYCHOLOGY[1]

A SPECIAL REVIEW
BY NINE PSYCHOLOGISTS[2]

The *Handbook of Experimental Psychology* was written to meet the "need for a technical survey which would systematize, digest and appraise the mid-century state of experimental psychology." More than four years in preparation, this monumental volume represents a distinguished achievement in scientific commentary. Whether measured in terms of comprehensiveness of content, quality of scientific thought, editorial skill, lucidity of discourse, length, number of illustrations, or *weight*—the *Handbook* sets new standards in psychological literature. All kinds of psychologists can feel proud that this oldest branch of their subject rests at mid-century upon such solid foundations.

The task of reviewing such a book presents in miniature one of the major problems that Stevens faced in getting the *Handbook* written. To paraphrase a sentence from his Preface, no single psychologist can marshal the erudition [or the time] for an adequate review of experimental psychology's facts and findings. So the present review is a collective effort, with eight reviewers dividing up the thirty-six chapters and the editor of this JOURNAL offering general comments. This procedure does not produce an integrated, balanced evaluation of the book, an outcome which partly reflects the fact that the writings of thirty-four specialists did not produce an "integrated" account of the state of experimental psychology.

This judgment is not necessarily a criticism of the *Handbook*, since one might legitimately question the desirability of trying to achieve systematic "integration" in a technical handbook covering such a heterogeneous field as "experimental psychology." It is virtually a physical impossibility for thirty-four authors to achieve the unity and continuity of discourse expected of a single writer—even if all of them spoke a common scientific language for congruent theoretical purposes concerning a homogeneous range of phenomena. The critic has only the right to ask whether or not the editor's general conception of the field, his specification of technical areas, his choice of contributors, and his editorial guidance were such as to maximize the chances of producing a relatively complete and balanced appraisal.

[1] STEVENS, S. S. (Ed.) *Handbook of experimental psychology.* New York: Wiley, 1951. Pp. xi+1436. $15.00.

[2] J. E. ANDERSON, J. S. BROWN, W. R. GARNER, D. A. GRANT, W. E. KAPPAUF, J. L. KENNEDY, L. H. LANIER, L. A. RIGGS, and ELIOT STELLAR.

Stevens doesn't spend much time discussing his conception of the nature and scope of "experimental psychology," nor does he offer much in the way of justification of his choice of topics for the book. In the brief Preface he notes that "between the covers of a single volume it would be impossible to survey all of psychology that is truly experimental. . . . There would need to be arbitrary slicings made in a subject matter that is essentially continuous, and these slicings would need to coincide with the talents and interests of living experts." True enough, but this brief rationale of the necessity for arbitrariness does not adequately justify: (1) the omission of some such qualifying term as *general* before *experimental* in the title, in line with the reasoning behind Murchison's 1934 version of a handbook of "general experimental psychology"; (2) the inclusion of much experimental material that is not psychology (the physiological chapters); (3) the inclusion of psychological chapters that are not experimental ("Growth Curves" and "Selection").

I am not overly concerned about the semantics of the title, although this problem is one which will increasingly disturb experimentalists in such fields as personality dynamics and social psychology. The inclusion of the physiological chapters and the non-experimental psychology is more questionable, since the considerable amount of space devoted to these topics might better have gone to more adequate treatment of such a field as motivation. In a sense, a more accurate title for the book might have been: "Handbook *for* Experimental Psychologists," with a subtitle reading "With special attention to the interests of physiological psychologists." In this frame of reference the excellent physiological chapters are better justified. (In any case they will be highly valuable to all psychologists who want up-to-date information about the physiological mechanisms underlying behavior.) A poll would be necessary to determine whether such material is believed to be more useful to psychologists than equivalent space devoted to an expanded treatment of such topics as motivation and perception.

The reviewer would have preferred to see some reduction in the physiological material and omission of the "non-experimental psychology" in favor of at least two additional topics: (1) the general methodology of science, including the logic of experiment and the logical structure of experimental psychology; (2) experimental design and statistical analysis. Graduate students of psychology—and most of their teachers—need general methodological orientation quite as much as they need "an inventory of experimental psychology's facts and findings." With respect to the treatment of "experimental design and statistical

analysis" in a handbook, I recognize that this would be difficult and that specialized textbooks on this topic are available. At the same time, I think that a systematic survey of the issues and techniques in this field would be worth while here, with special reference to the experimental behavior of psychologists as revealed in samples of their investigation.

But these are perhaps relatively minor shortcomings, especially when measured against the magnitude of the undertaking and the general excellence of the over-all achievement. Let us then turn to the reviews of the several sections of the *Handbook*.

LYLE H. LANIER.

University of Illinois.

MEASUREMENT IN PSYCHOLOGY

Reviewed by DAVID A. GRANT

University of Wisconsin

Mathematics, Measurement, and Psychophysics: S. S. STEVENS. Ch. 1, 1–49.

Stevens deals with these topics with sharp and heady prose which only occasionally lapses into turgidity and rarely forces the harried reader to lexical indagation. All in all Stevens makes a difficult subject accessible and, indeed, fascinating. His mood is vigorous and constructive, and he achieves a happy marriage of pragmatism and mathematical formalism.

After a brief introduction, mathematical models are discussed, particularly those relating to numeration and the number domain. The topics chosen are fundamentally algebraical in character, but technical details are avoided and subjects such as attributes of relations, postulates of order, postulational methods in algebra, the concept of group, and the concept of invariance are outlined. (Curiously enough, in spite of the historical emphasis on additivity in connection with measurement, the modern theory of integration and other topics of mathematical analysis go unmentioned.) Stevens goes on to describe scales of measurement, and then enters into the general problems of psychophysics. Here (p. 31) he remarks that in a sense " . . . there is only one problem in psychology—and it is . . . the definition of the stimulus." In a sense he is right, but probably no two psychologists would agree in what sense. After outlining the problems of thresholds, determination of equality, order, equal intervals, equal ratios, and general stimulus rating, Stevens lists the psychophysical methods used to attack these

problems. Although he includes order of merit and quantal methods, he omits the method of equal appearing intervals. The chapter concludes with a section on probability and a section on measures and indicants or correlates.

Despite the technicality of the issues, they are discussed with few errors, and these seem to the reviewer to be piddling and debatable. For example, the statement, "Repeated tests can add to (the *probability* of an empirical assertion) but never clinch its certainty" (p. 3) implies an esoteric definition of probability. The erroneous statement that a " . . . group in mathematics is a set of *operations*," illustrates the reluctance of the non-mathematician to use the term *set* without saying set *of* something or other. The section on probability (pp. 44–47) is cluttered by contributions of Russell, Keynes, and Carnap, which, in this context, can only mislead the novice. Some of Stevens' morbid doubtings in this section would be markedly alleviated by therapy consisting of a careful perusal of various limit theorems of modern statistics plus a casual restatement of his purposes in measurement and psychophysics.

As indicated, the reader has little in the way of error of detail to contend with in Stevens' chapter. The reviewer, however, does have serious reservations about the general level of the treatment. In spite of the preface statement " . . . the handbook should address itself to the advanced scholar—to the graduate student who would use it as a textbook (vii)," the technical level is more suitable to the *Scientific American*. Stevens tells about his topics rather than presenting them per se. For example, he remarks, "We have seen how the suppression of the commutative law makes it possible to add quaternions to the number system" (p. 15). Actually we didn't see it at all; Stevens told us it was so. Also, no actual data are presented—the tone of the chapter is hortatory rather than expository. The fact is that the reader is too well insulated from the harsh but wonderful details by Stevens' smooth presentation. His attitude is shown in writing about the psychophysical methods when he remarks, "We shall not worry long, however, for methods have a way of being tedious" (p. 42). The reviewer feels that this is being too solicitous.

In general, however, the chapter is a pleasant introduction to topics that the psychologist of the future must master if he wishes to be other than a dilettante. Some students may be lulled into a false sense of security, thinking that they now know the essential relations of numbers and mathematics to the psychologist's operations. The ones who count will doubtless be stimulated to further investigation of these issues and by further study may master the technical details required as a sound basis for the broader integrations this area requires.

Physiological Mechanisms

Reviewed by John L. Kennedy
The Rand Corporation

Excitation and Conduction in the Neuron: Frank Brink, Jr. Ch. 2, 50–93.
Synaptic Mechanisms: Frank Brink, Jr. Ch. 3, 94–120.
Sensory Mechanisms: T. C. Ruch. Ch. 4, 121–153.
Motor Systems: T. C. Ruch. Ch. 5, 154–208.
Homeostasis: Edward W. Dempsey. Ch. 6, 209–235.

Brink. In Chapter 2, the author considers first the structure of the neuron as determined by microscopic examination, by x-ray diffraction methods, and by the use of polarized light, the latter particularly for the study of the myelin sheath.

After discussion of the electrical properties of the cell (the outside of the resting cell is positive in relation to the inside), Brink points out that "the optically defined structures of the axon are not identical with the structures detected by electrical methods."

Turning to the nerve impulse, Brink describes the classical phenomena of the nerve-muscle preparation and the action-potential, with particular attention to the work of Hodgkin, which demonstrated that "conduction is mediated by the electrical currents associated with the action potential."

The phenomenon of accommodation of nerve (the change in excitability during the flow of a direct current) is discussed in relation to Lorente de Nó's proposal that "a change in the potential difference across a nerve membrane will change the excitability."

Rhythmic excitatory processes in axons are discussed in considerable detail, ranging from attempts to explain the intervals between impulses in a train initiated by a constant stimulus to studies of rhythmic processes inherent in the axon structure. Brink concludes the chapter by a discussion of the source of energy in neurons. He suggests that evidence points to the existence of two nerve membrane structures.

In Chapter 3, Brink reviews the classical reflex phenomena of temporal summation, spatial summation, and facilitation as they reveal synaptic characteristics. There follows a section on transmission across the synapse. "The flow of action currents associated with presynaptic impulses, the release of chemical substances, the state of excitation of the postsynaptic cell determined by the prevailing environmental influences, all combine to determine whether or not transsynaptic excitation initiates an impulse." Both direct recording from single synapses and study of artificial synapses lead to the conclusion that the action current from the presynaptic fiber is the primary stimulus to the postsynaptic cell. However, a "transmitter" agent (chemical)

seems to be involved in neuromyal transmission and may be present in nervous system synapses.

After discussion of after-discharge, the chapter is concluded with a short survey of electroencephalography.

Ruch. The chapter on "Sensory Mechanisms" begins with a survey of neurological techniques for answering "how" questions concerning the sensory mechanisms. The author reviews ablation, transection, electrical stimulation, recording of evoked and spontaneous potentials from the CNS, operative techniques for studying the brain in a normal physiological condition, and the Horsley-Clark stereotaxic instrument for accurate destruction or electrode placement in brain tissue. He recommends the use of the higher primates as experimental subjects.

There follows a review of the histology of the brain, staining methods, degeneration, and studies of the axon cylinder.

In a major section on thalamus and cortex, Ruch points out that the thalamus and cortex are a functional unit, operating by means of reverberating circuits.

Evidence is presented for duplication of sensory areas in vision, audition, and somatic sensation. After summarizing information on the taste pathway, physiological neuronography (local application of strychnine to small areas of the brain), and corticothalamic connections, Ruch's final section is devoted to the neurophysiology of perception. Here the neural basis of two-point sensitivity on the skin is chosen as a model for the explanation of sensory discrimination in all sense modalities.

Ruch points out initially in Chapter 5 that a great deal has been learned since 1930 about the motor systems originating in the cerebral cortex which is not yet available in standard textbooks of physiology. In particular, he cites additional information concerning the functions of the extrapyramidal system and the reticular substance of the brain stem as involved in "voluntary" behavior.

Under the topic, "Physiology of the Motor Cortex," Ruch describes the temporal characteristics of cortical stimulation, including latency, facilitation, and instability of the motor point. There follows a section on reflexes in relation to the motor cortex, in which are described the phenomena of "extinction."

A major section on the pyramidal tract reviews the anatomy and physiology of the "primary" voluntary system and discusses the information supplied by studies of recovery of function after lesions of the motor cortex.

The extrapyramidal system is given similar extensive treatment, although the neuroanatomy of the system, as Ruch points out, is not completely known.

In conclusion, he states, "Finally, one may question the value of likening neural systems to servo systems. Such analogies, devoid of

mathematical treatment, are essentially allegorical, somewhat akin to Freudian psychology. Whether a mathematical treatment will lead to predictions capable of experimental verification remains to be seen. Otherwise, we have added little since 1826 to Bell's 'circle of nerves.'"

Dempsey. Chapter 6 is devoted to one of the great generalizations of physiology, namely, that organisms have regulatory mechanisms, the action of which maintains constancy of the internal environment. Cannon applied the name "homeostasis" to designate these steady states. Dempsey then illustrates the concept by reference to the control of temperature of mammals, acidity of the blood, and sugar concentration of the blood.

The autonomic nervous system and the endocrine glands are the two major systems of the body involved in regulation or control. Dempsey devotes a major section to the description of the major endocrine glands and their hormonal output. He reviews data on the testis and ovary, the adrenal cortex, the pancreas, the parathyroid glands, the thyroid, and the anterior pituitary.

The autonomic nervous system, or paravertebral system, is reviewed in terms of the three divisions, (1) the sympathetic system, (2) the parasympathetic system, and (3) the central nervous system representation. "Both sympathetic and parasympathetic activities are integrated into patterns in the hypothalamus." Dempsey mentions examples of hypothalamic regulation of carbohydrate and fat metabolism.

The concept of homeostasis has been considerably extended from the original formulation of Claude Bernard in 1859. Dempsey discusses these extensions into the description of steady state phenomena of somatic behavior, evolution, instincts, and intellectual functions.

Growth and Development

Reviewed by John E. Anderson
University of Minnesota

Mechanisms of Neural Maturation: R. W. Sperry. Ch. 7, 236–280.
Ontogenetic Development: Leonard Carmichael. Ch. 8, 281–303.
The Genetics of Behavior: Calvin S. Hall. Ch. 9, 304–329.
Growth Curves: Nathan W. Shock. Ch. 10, 330–346.
Phylogenetic Comparisons: Henry W. Nissen. Ch. 11, 347–386.

Even as large and as amazingly comprehensive a book as this suffers from the division of psychology into experimental psychology on the one hand, and all other psychologies on the other. Because "experimental" is not a field, but a method of securing data on problems in many psychological fields by imposing controls within and upon situations at particular times, the editor faced a difficult task in choosing content. Nowhere is this more evident than in the section on *Growth and Develop-*

ment which, except for a very short section on "Growth Curves," stops ontogeny at birth. This is true even of Carmichael, whose chapter title is "Ontogenetic Development." Perhaps there is little reason for a section on growth in a handbook on experimental psychology. If one is nevertheless included, there should be reasonable coverage of age changes after birth or, at least, some recognition of the relation of age to psychological phenomena by authors of other sections. With the exception of Hovland no such recognition appears. In the Index only twelve entries appear under *Age*. No reference is made in this or any section to Terman's study of gifted children or to other longitudinal studies of selected populations. Is following a population selected in terms of defined characteristics over a period of time, experimental or not? An obvious answer to this criticism is that the editor of a handbook cannot cover everything and that in other sources, age changes receive more adequate treatment. But psychologists have been criticized for their tendency to base generalizations about the entire human population which includes both young and old and dull and bright, upon restricted samples of white rats, docile sophomores, obedient soldiers, and neurotic women.

Sperry. In "Mechanisms of Neural Maturation," Sperry reviews experiments on many species. He supports the thesis that the inherent pattern of neuronal linkages is achieved by embryonic processes similar to those responsible for the grosser phases of neurogenesis and the development of other organ systems. These involve a subtle differentiation into a multitude of neuron types. The induction of cell differentiation through contact with other cells and tissues, which is so common throughout all growth, takes on a special significance in the nervous system because of the length of the cell. The refined specification of the neurons makes possible selective synaptic linkages formed on the basis of chemo-affinity. Although Sperry's almost exclusive concern with the maturation of the integrative structures of the nervous system leads him to avoid the problems of the maturation of behavior, he hopes that his repeated reference to the role of anatomical connections in shaping response patterns will not exaggerate the importance of this factor, since the dependence of orderly function upon neural architecture does not exclude dependence upon other factors. He cautions the reader by pointing out that the view that the qualitative specification of neurons is closely correlated with differences in their synaptic association is an inference. Since adult learning is unable to effect rearrangement of the basic structure, why should we assume that earlier learning can do what later learning cannot do?

Carmichael, in "Ontogenetic Development," presents a compact summary of the development of the fetus and of the sensory equipment which the organism possesses at birth. He directs his efforts to describing the zero point and the subsequent developmental changes of some

typical segments of human behavior. In the foetal period changes are primarily the result of maturation. Because external stimulation is so well controlled, almost all activity is the result of internal processes. While there is some evidence of reactivity and conditioning, there is no evidence of learning or complex voluntary behavior. He derives a law of anticipatory function ("functional capacity may be demonstrated experimentally in many action systems well before the time when the function is normally called upon to play an active and significant part in the vital economy of the organism") which can be extended to cover features of postnatal as well as prenatal behavior.

Hall. In "The Genetics of Behavior," Hall summarizes the experiments on selective breeding in animals in which marked genetic differences in maze learning, activity, and emotionality have appeared. Experiments on strain differences show evidence of genetic factors in audiogenic seizures, emotionality, aggressiveness, hypothesis formation, geotropism, temperature preferences, and speed of reaction. He concludes that heredity is a factor in all psychological traits. As an example of the effects produced by a single gene (black), he points to the tremendous differences between wild and albino strains of rats with regard to emotionality and aggressiveness. Much of the psychological research would be improved by a more rigorous isolation of characteristics by the use of inbred or homozygotic strains, in which genotypes are identical except for *X* and *Y* chromosomes. Analysis in terms of multiple factors has not produced particularly significant results, although it is quite possible that more psychogenetic research will reveal such factors. In general, our knowledge of psychogenetics is very limited, although it is of vital importance.

Shock, discussing "Growth Curves," defines growth, which is a fundamental property of living things, as irreversible changes in size or function with time, and excludes learning for which the parameter is the number of trials which may be related to time in an arbitrary manner. A growth curve is a plot of size or function against time. A rational equation is presented from which several of the empirical equations developed for growth curves can be derived. For most physiological and psychological functions, except those in which time is the unit, the scientific possibilities of using growth curves are limited because of the inequality of units, even though they are valuable for normative and practical purposes. Estimates of growth rate in terms of various increments are discussed. While the material presented is very stimulating, it is narrow in scope and leaves out much developmental material of psychological significance.

Nissen's chapter on "Phylogenetic Comparisons" is an extraordinarily effective summary of a very large area. By limiting himself to the principles which emerge from phylogenetic comparisons, he stresses the systematic and theoretical approach. After some summary of methods, he turns to motivation and successively discusses homeostasis,

the motivation of perception and play, the innate determination of directive factors, and the modification of motivation by experience. He then considers cognition, and moves on to learning and concept formation. In general, differences in wants among various species are trifling in comparison with the differences in the means of satisfying those wants. The significant axes of evolution appear in cognitive rather than motivational behavior. They are found (*a*) in sensory differentiation which increases the possibilities of perceptual organization; (*b*) in the emergence of symbolization, a new instrumentality of particular significance among human beings, and (*c*) in the emergence of one-trial learning as a result of (*a*) and (*b*), which among the highest primates replaces the many-trial learnings characteristic of lower forms.

If a chapter similar to Nissen's, which summarized in a broad way the principles and facts of ontogeny from birth to death, had been included, this section would have been well rounded out. But whatever one may say, the *Handbook* is an outstanding contribution that will be of great usefulness to psychologists. Not only have the writers taken their responsibilities conscientiously, but they also have shown insight in bringing such excellent material together in such a stimulating way. The result is more than impressive; it is almost overwhelming. One wishes that as carefully prepared and as complete a handbook were available for other areas of psychology.

MOTIVATION I

Reviewed by ELIOT STELLAR
The Johns Hopkins University

Instinctive Behavior: Reproductive Activities: FRANK A. BEACH. Ch. 12, 387–434.
Emotion: DONALD B. LINDSLEY. Ch. 14, 473–516.

Modern treatment of the problem of motivation typically divides the topic into two parts: the biological drives and the acquired or learned drives. As often as not, the emotions are included as a special instance of motivated behavior. The section in the *Handbook* on motivation makes the more inclusive, three-way division of the problem and is an excellent job with one exception. Unfortunately, the treatment of the biological drives is limited to a discussion of the broad area of reproductive behavior. Nowhere in this comprehensive review of experimental psychology is there any treatment of thirst, hunger, or specific hungers, let alone the experimental literature on such behavior as homing, migration, and nesting.

With the space allotted to the topic of motivation limited, there can be no argument with the decision that reproductive behavior be given priority over the other biological drives, for we know the most about it, and it is of great importance at the human as well as the animal level.

But it is a real question whether the space that could be devoted to motivation should be sacrificed for some of the topics included in other sections.

Aside from these omissions, which are no fault of the authors, the section on motivation is superbly done. The chapters are scholarly works, they are authoritative, and they offer excellent coverage of the relevant literature. In this section of the review only Beach's chapter on reproductive behavior and Lindsley's chapter on emotion will be considered. Miller's chapter on learnable drives and rewards will be treated separately in the next section.

Beach. In the chapter on reproductive behavior, Beach accomplishes the herculean task of presenting concisely the tremendous literature available on the neurophysiology, biochemistry, and psychology of sexual behavior, parental behavior, and filial behavior. The chapter had to be tightly organized and tersely written, and these things Beach did. The first half of the chapter is devoted to sexual behavior, and the last half to parental and filial behavior. Each section is divided into four parts: sensory factors, neural factors, blood chemistry, and the role of experience. Within each of these parts, there are subsections in which the relevant variables influencing behavior are discussed. And finally in the discussion of each variable, the experimental data are presented, wherever possible, species by species up the phylogenetic scale.

Because of the mass of data presented, the chapter will not be easy reading for students. But it should be profitable reading, for Beach manages to tie most of the data together from time to time throughout the chapter in a series of clear, general conclusions. There is not space to list the conclusions here, but it is clear from them that reproductive behavior is controlled by a variety of learned and unlearned mechanisms, sensory, endocrine, and neural. Furthermore, in the course of evolution, endocrine control becomes less important and higher forebrain mechanisms more important. Little systematic change occurs in the basic sensory control, but in the higher organisms learning assumes more and more importance.

In view of the complex physiological and psychological control of reproductive behavior, Beach seriously questions whether the term "instinct" is appropriate. He recommends dropping the term from the scientific vocabulary, and it is too bad that it will be perpetuated for another generation by getting into the title of the chapter.

Lindsley. In the chapter on emotion, Lindsley offers a broad survey of the physiological basis of emotional behavior. He deliberately, and justifiably, omits any consideration of the subjective aspects of emotion on the grounds that we lack reliable data.

The main treatment of emotion is divided into three parts. The first section is devoted to a discussion of sixteen measurable bodily changes that accompany emotion, including such standbys as GSR,

EKG, blood pressure, and respiration. The second section covers the autonomic nervous system, and the role in emotion of such humoral factors as adrenin, insulin, sympathins, and acetylcholine. In the third section on the central neural organization of emotion, evidence is brought together from studies of experimental lesions, electrical stimulation, and the analysis of EEG records.

Out of the discussion of the central mechanisms in emotion comes Lindsley's outstanding contribution, the *activation theory* of emotion. Briefly put, the theory rests on two facts: (1) in emotion, there is a substitution of small, fast brain waves for the larger, slower alpha rhythm and (2) such substitution of fast for slow waves can be produced by stimulation of the reticular formation in the brain stem, and conversely, lesions of the reticular formation or in its connections with the hypothalamus result in a persistent pattern of large, slow waves. At one extreme, Lindsley places sleep, where the brain waves are largest and slowest and activity and excitement are at a minimum. Next is relaxed wakefulness where alpha rhythms predominate. Then there is sudden stimulation or mild surprise where there is transient substitution of fast for slow waves. Finally, there are the states of mild emotion, apprehension, and anxiety where there is more and more persistent appearance of the small, fast waves in place of the slower alphas.

The achievement of this theory is that it is based on good neuroanatomical facts. Furthermore, not only does it have the advantages of the Cannon-Bard theory, but it is more detailed and also serves to bring emotional behavior into a logical continuum from sleep, through alert wakefulness, to chronic anxiety and highly excited states.

Motivation II

Reviewed by Judson S. Brown
State University of Iowa

Learnable Drives and Rewards: Neal E. Miller. Ch. 13, 435–472.

Readers primarily interested in the construct of motivation as a behavior determinant may be somewhat disappointed by the limited treatment according this topic in the *Handbook of Experimental Psychology*. According to the Table of Contents, Chapters 12, 13, and 14, which deal, respectively, with instinctive sexual behavior, learnable drives and rewards, and emotion, comprise the section on motivation. But neither the first nor the third of these appears to contain a single explicit reference to motivation *qua* motivation—at least none of sufficient prominence to have justified its inclusion in the index under either *motivation* or *drive*. Most conspicuous by its absence is a chapter on primary drives and their bodily bases. The original plans for the *Handbook* apparently called for such a chapter, but its completion was prevented by the author's illness. Chapter 13 does deal explicitly with

secondary motives. But it alone can scarcely be expected to fill the hiatus created by the absence of an adequate formal treatment of primary drives.

In Miller's discussion of learnable drives and rewards, attention is focused primarily upon (1) the acquired drive of fear, and upon (2) acquired rewards and drives based upon food and hunger. Additional material on other drives as sources of secondary motivations, on complex social motives, and on general theoretical problems is included in minor supplementary sections. Throughout the chapter, the framework of Hull's learning theory and of the ancillary Miller-Dollard version are revealed as effective integrating schema.

Without question, the initial section on fear comprises the most exhaustive treatment of any particular secondary drive yet to appear. And it is only natural that here the theoretical and experimental work of Miller and Mowrer, which constitutes the bulk of the relevant material, should be given a place of central importance. According to these authors, fear can be most usefully conceptualized as a conditionable response possessing functional properties akin to those of primary drives. Such being the case, the principles governing its acquisition, maintenance, elimination, and motivating functions, should, in essence, be the same as the principles appropriate to other responses and other drives. Broadly viewed, the section on fear is a résumé of attempts to determine the variables affecting fear and the applicability to fear of existing principles of learning and motivation. This is not to imply, however, that the chapter is nothing more than a digest of experiments. The entire work is marked by critical acumen of high degree and is rich in stimulating hypotheses and suggestions for future research.

The second principal section, on learned rewards and drives based on food and hunger, begins with a straightforward review of experiments on secondary reinforcement. Here the presentation is nicely facilitated by a tabular summary of the important aspects of all the relevant studies. Where the treatment shifts, however, to the problem of learned *drives* based on hunger it becomes rather nebulous. The important question at this point is whether hunger, as well as fear, can be conditioned. Unfortunately, no analysis is presented of the conditions under which such a secondary hunger drive might be acquired, how its presence might be unequivocally demonstrated, or how it might be clearly distinguished from the incentive functions of tokens. Because these admittedly vexing problems have been skirted, the casual reader may get the erroneous impression that convincing demonstrations of the conditionability of hunger have indeed been carried out. Actually, none of the studies purporting to demonstrate this phenomenon has been successful in ruling out a number of plausible alternative interpretations.

In spite of minor defects such as this, Miller's chapter most certainly constitutes a significant contribution to the psychology of learning.

It will be widely accepted as the definitive work on acquired drives and rewards for a considerable period to come.

Learning and Adjustment

Reviewed by David A. Grant
University of Wisconsin

Methods and Procedures in the Study of Learning: E. R. Hilgard. Ch. 15, 517–567.
Animal Studies of Learning: W. J. Brogden. Ch. 16, 568–612.
Human Learning and Retention: Carl I. Hovland. Ch. 17, 613–689.
Theoretical Interpretations of Learning: Kenneth W. Spence. Ch. 18, 690–729.
Cognitive Processes: Robert Leeper. Ch. 19, 730–757.
The Psychophysiology of Learning: Clifford T. Morgan. Ch. 20, 758–788.
Speech and Language: George A. Miller. Ch. 21, 789–810.

Hilgard presents the methods and procedures for studying learning in a clear, straightforward, and unpretentious style. The tone is thoroughly empirical. In defining learning, Hilgard remarks (p. 518), "The experiments themselves define the field ostensively."

After brief descriptions of the essentials of a dozen or so learning experiments, Hilgard goes into the procedures to be followed in classical conditioning. Instructions, control tests, optimal intervals between stimuli and trials, and arrangements for conditioned discrimination are outlined along with the types of conditioning scores which result. Similar material is given for instrumental conditioning. Procedures in maze and discrimination experiments are then discussed, and sample tables are given for rat age-weight norms, balanced trial orders, and statistical criteria for learning. The reliability of various scores and types of apparatus is summarized. The section on motor skills contains little material on the methodology of motor skills research per se. This reflects adversely on the research in the area rather than on Hilgard's treatment.

Hilgard is at his best in discussing memorization and retention. This section should go far towards preparing the novice to plan and carry out an experiment in the field, and can even be read with profit by the experienced researcher. About ten pages of tables of calibrated nonsense syllables and adjectives are presented. Here a good deal of Melton's unpublished material is made generally available, a welcome addition to the literature. Valuable suggestions on the definition of overlearning and the finer points of retention scores such as those given here could well serve as examples of statistical and experimental horse sense which might well be applied by all of us in setting up scores and indices in our own areas of research. For example, Hilgard points out

(p. 556) that if the criterion for original learning is a certain percentage of correct responses, the base for 100 per cent recall must be per cent correct on a control trial following the criterion trial.

In discussing learning curves as analytic devices, the author outlines Vincent curve procedures and the fitting of rational and empirical curves. The author's feelings for priority combine with lack of space and the date the chapter was completed to deprive the reader of some excellent examples of rational curves. Hilgard winds up in a section on "Next Steps" by favoring more precise methodology, more appropriate conceptual formulations, more miniature systems (and recognition that current systems are miniature) and more fresh starts with naturalistic studies. Who would quarrel with these?

The reviewer detected no flagrant errors in this chapter. The limitations characterize the field as a whole. There are a few minor points, however: (*a*) our data do not favor first establishing the positive CR in eyelid conditioned discrimination (p. 528); (*b*) matching *S*s may introduce some experimental economy (p. 536), but unless correlations are *exceptionally* high the results are disappointing, not to say useless; (*c*) complex rotations of procedures (p. 537) need not preclude complete statistical analysis; (*d*) much of the work on proactive inhibition has centered on the loss of retention in the material learned after the prior learning rather than in the difficulty of the test learning itself (p. 557). All in all, Hilgard's chapter accomplishes its mission; it should be most useful both to graduate student and mature researcher.

Brogden packs a tremendous amount of detail into his chapter on animal learning. The number of factual items per page and the precision of diction make a second reading of most sections rewarding.

In preliminary remarks, Brogden points out that his preference is for functional experiments in which the dependent variable is measured at several values of the independent variable rather than for the so-called crucial experiment with one experimental and one control group. The latter are cited, for the most part, in a supplementary way only.

Conditioning, discrimination learning, and serial or maze learning are each treated in terms of acquisition, transfer, and retention. The terminology is similar to that of Hilgard and Marquis' *Conditioning and Learning* but the organization will inevitably excite some adverse comment; certainly it means more work for the reader. It seems to the positivistically inclined reviewer that there is considerable merit in some aspects of the classification, and that the facts of learning are not so well understood but what an occasional reorientation is helpful.

The general accuracy is high, and the organization points up several lacunae in our data. The reviewer would never classify pseudo-CR's as forms of backward conditioning (p. 580), and there appears to be at least an ambiguity where it is stated that spontaneous recovery is proportional to the extent of experimental extinction (p. 589). The loudest screams will come from those who resent Brogden's ruthless

pruning out of the "crucial" experiments and theoretical interpretations with which the animal literature abounds. At least he is impartial, giving no interpretation of the remarkable identity of the coefficients of the Gompertz curves fitted to his own data (p. 584). Then too it is a notorious fact that the today's crucial experiment is generally found to be irrelevant tomorrow, whereas dimensional studies have a somewhat greater life expectancy. The student will be able to find a wealth of facts in the chapter and may be stimulated to try to fill some of the gaps in our knowledge that Brogden has pointed out.

Hovland. Like Brogden, Hovland's treatment of human learning and retention has a strong empirical flavor. This must be done " . . . since analytical treatment of theoretical variables often requires the type of simplification and control attainable only with animal experimentation" (p. 613). Once adopted, this attitude permeates the whole chapter and results in the deletion of some theoretical interpretation even where available.

Hovland first gives brief summaries of conditioning, verbal learning, and motor learning. He next goes on to deal with motivation, individual differences, efficient methods, distribution, whole-part features, recitation, retention, reminiscence, transfer of training, causes of forgetting, and mathematical formulations of learning. For the most part these treatments are excellent, accurately presenting vast amounts of data. On the other hand the *causes of forgetting* section seems to be included only for completeness or because McGeoch used such a heading.

Hovland's chapter is carefully done and generally critical. Unlike some of the chapters it was apparently not supported by a military contract nor revised from sections of the author's other publications. The reviewer feels that a little too much weight was given to priority so that in some instances more adequate later studies weren't cited. For example, Underwood's work might have appeared more frequently in the text. Occasionally necessary warnings are omitted; e. g., the probable influence of sequence of procedures in Luh's study of retention (p. 647). Another general comment: failure of later studies to obtain statistically significant results is usually noted, but the fact that earlier studies included no statistical analysis is omitted. Finally, the reviewer is baffled as to why learning-curve equations derived from differential equations are classed as empirical, while Hull's positive growth function is classed as rational (p. 678).

In general, however, the student who studies Hovland's chapter will get an excellent view of the research on human learning. He will note a good deal of progress in detail since Hunter's chapter in the old *Handbook* and McGeoch's *Psychology of Human Learning*, but few really significant changes.

Spence's treatment of learning theory is a lucid and able presentation of the *S-R* approach to which he has made so many contributions.

After describing the *S-S* vs. *S-R* issue and the reinforcement issue as the issues dividing learning theorists, Spence gives an elegantly concise outline of Hullian behavior theory in relation to classical and instrumental conditioning. Contiguity and two-factor theories are also outlined more briefly. The relative difference in space allocation can be defended on the basis of the more complete elaboration of "behavior theory," but Skinner's formulation gets only scant attention, and the treatment of Hullian theory is understandably more up to date than that accorded to its rivals. The reinforcement-contiguity issue is explored generally. Pro-reinforcement psychologists will probably agree with the arguments of the chapter, but the "antis" will not. Spence has been through all this before, and it is available elsewhere. He does his stint ably, but he feels that " . . . too much attention has been given to this issue of *S-S* versus *S-R* . . . what is needed most . . . is the continued and persistent (pursuit) of the type of integration of theory and experiment that will lead to the discovery and formulation of laws of learning in terms of environmental and behavioral variables" (p. 725). With this few psychologists would quarrel.

Leeper. In the chapter on "Cognitive Processes," Leeper attempts to deal with a difficult topic. Most psychologists have had trouble with the cognitive processes, and Leeper is no exception. There is a paucity of sound experimental literature on the topic, but only a small portion is cited. Leeper indulges first in a lengthy discussion of the introspective attacks on cognition and the imageless thought controversy. This warmed-over hash will scarcely be appetizing to the modern psychologist, but to be impartial, Leeper goes on to misuse the term "introspection" where "verbal report" would be correct, which will doubtless annoy some who care about these matters. Inductive concept formation, deductive concept formation, and inventive concept formation are outlined in terms of one or two experiments each, and then a brief listing of factors favorable to and unfavorable to problem solving are described with cited examples.

A good introduction for the graduate student to this topic is admittedly hard to achieve, but the reviewer feels that the present chapter may be unnecessarily inadequate.

Morgan. Unlike the other authors, Morgan is able to cover the psychophysiology of learning in a relatively exhaustive fashion. Although there are few recent references (only ten since 1945) most of the relevant literature is cited. The neural locus of conditioning is first discussed with the conclusion that the process takes place " . . . some place in front of the final common motor path from the cortex" (p. 761). Spinal conditioning is presented with some judicious doubts, and then Morgan departs from strictly learning material to describe the effects of cortical lesions on various sensory discriminatory capacities. Effects of lesions on locomotor and manipulative learning are contrasted, but again the learning aspect seems secondary in the

studies cited. "Equipotentiality" is modified into "neural equivalence" in describing recovery of motor and sensory functions, and this is followed by an excellent section entitled "Selection and Variability" which deals with selective response of the organism to multiple cues and the shifting of response from one cue to others. The final section deals with the localization of memory functions in the brain, in which the emphasis is upon the mysterious prefrontal areas. The chapter ends with a brief description of aphasia, agnosia, and apraxia.

Morgan weaves a coherent thread through a most complex pattern of data. When possible he arrives at a definite principle or conclusion; indeed, sometimes he fails to reserve doubts when he should; e.g., when he states flatly that prefrontal monkeys cannot solve complex problems (p. 783). Perhaps some will feel his account is over-Flesched, but the novice will appreciate the clarity of the presentation. The reviewer feels more should have been given on methodology and some mention should have been made of the special criteria by which to judge extirpation studies; e.g., effect of method of making lesion, likely invasion of underlying centers and pathways, provision for complete operative recovery, and nature of histological appraisals.

Unfortunately the splendid studies of Woolsey and his co-workers on secondary, sensory, and motor areas apparently appeared too late for digestion of their radical implications for earlier extirpative work. All told, Morgan has done a thoroughly competent job, and the novice will have been guided a long way by this cautious summary of the area.

Miller essays a chapter on "Speech and Language," a fine addition to the *Handbook*. In the limited space, the treatment is restricted "to the subject of verbal context" (p. 789). By verbal context is meant " . . . the communicative acts which precede and follow the verbal response under consideration" (p. 790). The context is thus a time series, and Miller proceeds to apply Shannon's procedure to approximate English text by statistical means. He next considers the capacity of the speech mechanism. By a series of assumptions, Miller establishes 67 bits/sec. as the theoretical physiological limit, with a reduction to 32 bits/sec. imposed by the phonetic structure of English. Then considering limitations on vocabulary and the dependency of successive words, the figure is reduced to 7 to 9 bits/sec. Next, the associative structure of an individual's language is discussed in terms of techniques used to investigate relations between words. Readability is granted a brief section, and the chapter ends, rather weakly, on the topic of talking and thinking.

This is not a typical *Handbook* chapter, bulging with data and citations, but within the context of the author's aims one might criticize omission of the requirement that Shannon's time series be stationary—an important limitation. In the discussion of "Talking and Learning" one finds no comment on symbolic processes in general nor, save for

brief mention of Zipf, any consideration of the function or purpose of language as a factor in determining its structure.

The chapter is a welcome innovation. Future *Handbooks* will doubtless see its coverage extended.

Sensory Processes I: Vision

Reviewed by Lorrin A. Riggs
Brown University

Basic Correlates of the Visual Stimulus: D. B. Judd. Ch. 22, 811–867.
Visual Perception: C. H. Graham. Ch. 23, 868–920.
The Psychophysiology of Vision: S. Howard Bartley. Ch. 24, 921–984.

Judd has summarized the terms, scales, units, and computational methods currently accepted for specifying light as a stimulus for vision. The last part of the chapter is a treatment of color vision, organized around the topics of color theories, the perception of surfaces, standard chromaticity diagrams, and the specification of color by the Munsell system.

This chapter contains much that is new in psychological literature. Particularly valuable reference material includes the following: (1) A clear description of the ICI system of colorimetry. (2) Tables and conversion factors for the baffling array of radiometric and photometric units which are still in common use. (3) A diagram nicely illustrating the gradual course of the Purkinje shift from scotopic to photopic levels of adaptation. (4) A table summarizing important theories of color vision, including two "zone" or "stage" theories which antedate Granit in their use of concepts similar to dominators and modulators. (5) A glossary of important terms. Judd's chapter is perhaps definitive to the extent that the modern terms illuminance (in place of illumination), luminance (photometric brightness), troland (photon), and luminosity (visibility) will now find acceptance among psychologists. Certainly the new terms are to be recommended as less ambiguous than the old. The use of decibel notation for visual intensity has less to recommend it, however, particularly as it has arbitrarily been related to the threshold of extra-foveal cone receptors at 560 mμ. This leads to the use of negative *db* units for rods, and emphasizes the point that human vision, unlike hearing, has various ranges of sensitivity depending upon the region stimulated, the level of adaptation, and the type of receptor involved.

Judd has given a disproportionate amount of space to the "engineering" details of color specification, particularly in relation to the Munsell notation. It might have been more useful to devote this space to a description of instruments and methods currently available for photometry and colorimetry.

Graham's chapter is distinguished as much by its point of view as

by its coverage of the literature on visual perception. Some of the main points may be paraphrased as follows: (1) Experiments in visual perception are best set up to use psychophysical methods in which a limited number of responses are permissible. (2) These responses are made to stimulus operations such as the presentation of objects, words, or energies. (3) The relevant characteristics of the stimulus must be judged in terms of adequate psychological theory. Merely physical descriptions of stimulating conditions may often be of less consequence than psychological ones, as in the situation where the subject gives the same response to a whisper as to a shout. (4) The basic data consist of response frequencies as they relate to variations in the relevant characteristics of the stimulus. (5) Such data represent discriminations which traditionally have been subsumed under subjective terms such as "perceive," "see," "detect," "appear," etc. (6) The terms "sensation" and "perception" cannot be distinguished on the basis of psychological or psychophysical functions. The distinction seems merely to be that, historically, "sensation" has been used where some relevant sense organ theory exists, while "perception" has been used in the absence of such theory. (7) Current theories of perception are open to criticism on the basis that they do not specify operations by which they may be tested. (8) An adequate theory of visual perception must relate the responses of the individual to aspects of the stimulus, frequency and time of presentation, and conditions of the organism. (9) Such a theory may well consider physiological as well as purely psychological concepts and terms; in particular, vision theories such as those of Hecht and Granit may profitably be used, as well as behavior theories such as those of Hull and Skinner.

Of particular interest in Graham's chapter are quantitative treatments, within the framework just outlined, of the topics of monocular movement parallax, stereoscopic vision, span of perception and figural after effects. For each of these topics there is a discussion of the significant variables and a summary of recent experimental work, much of it contributed by Graham and his associates. Other topics, the existing data for which are not so readily fitted into this framework, are more briefly discussed. These topics include the cues for visual space perception, the discrimination of size, the discrimination of shapes, binocular rivalry, illusions, real and apparent movement, fluctuations of perception, and the recent experiments in which perception is importantly affected by the condition of the organism (needs, set, anxiety, etc.).

Bartley's chapter is a readable and reasonably comprehensive treatment of visual psychophysiology. The material is divided into sections on (1) optical, (2) anatomical, (3) photochemical, (4) neural, and (5) oculomotor mechanisms. To this reviewer it appears that Bartley has achieved his wide coverage at the expense of failing to come to grips with any of the problems which he has chosen to discuss. The analysis of relevant variables, so prominent a feature of the chap-

ters by Judd and Graham, seems generally to be replaced in the present chapter by a mere recital of certain facts arbitrarily selected from the massive literature on each topic. The topic of dark adaptation is treated, for example, without reference to visual purple, vitamin A, or the dynamics of the visual cycle worked out by Wald, Morton, and others.

This chapter is poorly documented. There is a description of the findings of Hartline on single optic nerve fiber discharges in the vertebrate eye, but no mention of his name except in the last paragraph on the "receptive field." Figures 48, 49, and 51, attributed to Hartline, appear to be crude pen drawings of the original and classic oscillographic records. In a section on the electroretinogram there is a detailed description of the analysis, actually made by Granit, of component processes PI, PII, and PIII. There are also specific descriptions of recent experiments on the human electroretinogram by Adrian, Riggs, Johnson, and others. Yet the whole section on the electroretinogram contains but a single reference, namely one to an article by Bartley entitled, "Some factors in brightness discrimination." There is no mention of Miles in the section on red goggles for dark adaptation.

Taken together, the three chapters on vision are found to cover most of the important material. Judd has given us a unique compilation, in the best handbook tradition, of reference material on light as a stimulus for vision. Graham has advanced a new point of view on visual perception and has demonstrated its fruitful application to certain of the problems in that area. Bartley has presented sample data on a wide variety of topics in the broad field of psychophysiology. Nowhere in the three chapters does one find an adequate description of the eye and the visual pathways. An optical diagram of the eye, showing the nodal point, center of rotation, pupil, macula, and blind spot might well have been used to clarify the discussions of such topics as retinal image formation, monocular movement parallax and the Stiles-Crawford effect. A few errors may be mentioned. On page 817, Formula (3) should evidently read $E(\text{lux}) = .929\ Td^2$ (square centimeters) $\times B$ (foot-lamberts). On page 892, in the legend for Fig. 21, the word illumination is used to refer to photometric brightness or luminance in millilamberts. On page 945, the equation in parentheses should read, (1 foot-lambert = 1.076 millilamberts).

Sensory Processes II: Audition

Reviewed by W. R. Garner
The Johns Hopkins University

Basic Correlates of the Auditory Stimulus: J. C. R. Licklider. Ch. 25, 985–1039.

The Perception of Speech: J. C. R. LICKLIDER AND GEORGE A. MILLER. Ch. 26, 1040–1074.

The Mechanical Properties of the Ear: GEORG VON BÉKÉSY AND WALTER A. ROSENBLITH. Ch. 27, 1075–1115.

Psychophysiology of Hearing and Deafness: HALLOWELL DAVIS. Ch. 28, 1116–1142.

These four chapters on Hearing present a nearly complete treatment of their subject—probably a more complete treatment than is afforded any other topic in the *Handbook.* This completeness of treatment is not just a question of allotting more space to the subject. Although four chapters may seem like too many for "Hearing," the number of pages used is actually less than for "Vision." The completeness comes about rather because of the very systematic nature of the treatment, because there is practically no overlap between the various chapters, and because every possible subtopic is carefully fitted into its place. In fact, to the reviewer these four chapters are a good illustration of the best that can be done in handbooks. There is good consistency, even though five different authors are involved, and there is clear evidence of a carefully thought out plan for the entire set of chapters. A great deal of the cohesion which these chapters show is undoubtedly due to the physical and mental proximity of the authors. But whatever the cause, they can well be used as examples of good handbook organization and writing.

It is always possible to argue about what should be contained in a handbook, and seldom will two different psychologists (or anybody else) agree exactly. There are places in these chapters where one could argue that the material is not treated in a manner best suited to handbook writing, but it is doubtful that one will argue that these chapters are not well done if the obvious intent of the chapters is accepted.

Licklider's chapter on the correlates of the auditory stimulus is the longest of the four, and it covers what might be called the psychophysics of auditory sensation. After several pages on considerations of the nature of the sound stimulus and various ways in which the stimulus can be described, there follow discussions of absolute and difference thresholds, the various dimensions of tonal experience, masking and fatigue, beats, harmonics and combination tones, subjective attributes of complex sounds, temporal effects in hearing, sound localization and other binaural effects. There is an appropriately small space allotted to beats, difference tones, etc., and more space to such problems as masking, auditory fatigue, and critical bands. It is pleasant to see this distribution, because psychology texts, particularly the introductions, have for so long depended heavily on the former problems, and have practically ignored the latter. There are 43 illustrations, and the only real criticism about this chapter concerns them. The illustrations were constructed to show the nature of various relations very well,

but they do this at the expense of having graphs from which values can be read. For example, three of the graphs are three-dimensional portrayals. Such graphs are very pretty, but it is a shame not to be able to read actual values for equal loudness contours and loudness in sones.

Licklider teams up with *Miller* for the chapter on the perception of speech. The handling of the material is fairly conventional, but of course this conventionality is not too familiar to psychologists. After an initial discussion of ways of physically measuring and describing speech, there is a section on methods of measuring the intelligibility of speech. Then follows a discussion of various factors which affect intelligibility, including several types of distortion. A short section is used for a description of empirical methods of calculating the intelligibility of speech. The methods described stem mainly from the work of the Bell Telephone Laboratories, and the rather complicated procedures worked out at that Laboratory are presented in a highly simplified but comprehensible manner. The chapter concludes with a discussion of speech factors other than simple intelligibility.

As the authors point out, this material is truly psychological. Yet until recent years most of the work has been done by engineers and physicists. Of all the material in the area of hearing, this type has been least available, but is probably the most interesting to psychologists.

Békésy and *Rosenblith* present a thorough description of the mechanical properties of the ear, from the pinna to the basilar membrane. The chapter contains an excellent summary of the recent work (particularly by Békésy himself) on the mechanical dynamics of the inner ear. Little of this work has previously been published in psychological journals, and it will come as a real surprise to those psychologists who have not kept up with the hearing literature in other journals to realize how much is now known, and how highly specific theorizing has become as a result of this new knowledge.

Davis continues the description of the hearing mechanism into the nervous system. In this area there have also been some great advances due to the development of better electronic recording and stimulating techniques. The chapter discusses the simple neuroanatomy, and the experimental data available to show the nature of the neural process. In the latter part of the chapter there is a short discussion of deafness. This is the one topic which the reviewer feels should have had a more thorough treatment. There is quite a bit of material available on diagnostic techniques which are essentially psychological. Perhaps the *Handbook* is not the place for a discussion of clinical techniques, but the psychophysics of deafness forms the basis of these techniques and a discussion of it certainly would have been in order.

All in all, however, any criticisms of these four chapters are minor compared to the over-all quality. An occasional omission is probably necessary, and it would be difficult to say what should have been omitted in order to include the material on deafness.

SENSORY PROCESSES III

Reviewed by W. R. GARNER AND ELIOT STELLAR
The Johns Hopkins University

Taste and Smell: CARL PFAFFMANN. Ch. 29, 1143–1171.
Somesthesis: WILLIAM LEROY JENKINS. Ch. 30, 1172–1190.
Vestibular Functions: G. R. WENDT. Ch. 31, 1191–1223.
Time Perception: HERBERT WOODROW. Ch. 32, 1224–1236.

It is a matter of historical practice in psychology to call the chemical senses and somesthesis the minor senses, and then to devote as little space and time to them as possible. And when the vestibular functions are treated as one of the senses, they get similar treatment. The present *Handbook* is no exception. Such scant coverage is perhaps justified in terms of their subjective unimportance and our ignorance about them. For all three of these topics there is even difficulty in defining the physical stimulus, as the authors point out. We are not sure what receptors are stimulated, or how they are stimulated. Consequently, quantification is extremely difficult. In his chapter Wendt even raises the question of whether there is a vestibular sense at all.

Despite these limitations imposed by our relative ignorance, these three chapters are up to date, scholarly, and lucid, just as they should be in a handbook of Biblical proportions. The only criticisms one can level against these chapters is that the limitations of space forced the authors to omit or give minimal coverage of important topics and areas of research.

Pfaffmann takes up the chemical senses, taste and smell, in that order. In each case he describes the relevant anatomy, the nature of subjective qualities, the theory of stimulation, and the factors influencing the sensitivity of the organism. Most significant perhaps is his excellent discussion of the conflicting and confusing theories of the nature of the physical stimulus and its mechanism of action. He was, however, forced to cover the very instructive material on the chemical senses of invertebrates by reference to an excellent review of the topic.

Jenkins divides the material on somesthesis into cutaneous sensitivity, subcutaneous sensitivity, kinesthesis, and organic sensitivity. Most space is devoted to the skin senses, and here the author goes a long way toward clearing up the classical confusion. He points out that there may actually be three touch senses, two pains, and perhaps two warms and two colds. Even more important, he disposes once and for all of the simple notion that there is point-to-point correspondence between sensory spots on the skin and specific receptors within the skin. His greatest omission is of a significant treatment of the role of the central nervous system in somesthesis. Although some of this material is covered in Chapter IV, the present chapter is seriously unbalanced by failure to mention the experimental and clinical literature on central

mechanisms operating in somesthesis, particularly the classical work of Head and his associates.

Wendt's chapter on vestibular functions, after a few pages of anatomical description, considers vestibular control of the eyes, of the neck muscles, of the trunk and limbs, and of motion sickness. The end of the chapter is concerned with the perception of movement and the modifiability of the vestibular system. Usually vestibular functions are discussed under the heading of adjustive processes (as in the Murchison *Handbook*), or the more specialized topic of reflexes. And usually a physiologist writes the chapter and spends much more time on such topics as the role of head movements in postural reflexes. The fact that this chapter was put in the section on sensory processes and the fact that it was written by a psychologist have combined to give the material a much better psychological orientation than is usual.

In all three of these chapters the omissions were undoubtedly deliberate. The authors had to decide what to put in and what to leave out. We cannot seriously disagree with their decisions, for we cannot decide what should have been left out in order to include other material.

Woodrow. The last chapter in this section is a short one on time perception. It is necessarily short, since there is after all little that can be said about time perception with any degree of generality. Most of the chapter is taken up with a discussion of the estimation of short time intervals, and of the effects of defining the time interval in different ways. There are also short sections on the experience of unitary duration (i.e., temporal span of attention), rhythm, and absolute judgments of time. Although the material is interestingly presented, it is doubtful that even this much space should have been allotted to the topic. The chapter has much more the tone of a congenial discourse than do others in the book, because there is much less need to crowd facts in. Probably the most important role this chapter can play is to stimulate further research in the area.

Human Performance

Reviewed by William E. Kappauf
University of Illinois

Selection: Harold P. Bechtoldt. Ch. 33, 1237–1266.
Training: Dael Wolfle. Ch. 34, 1267–1286.
Engineering Psychology and Equipment Design: Paul M. Fitts. Ch. 35, 1287–1340.
Work and Motor Performance: Robert H. Seashore. Ch. 36, 1341–1362.

These four chapters describe research techniques and research evidence which the experimental psychologist employs when dealing

with certain applied problems. As such, these are new chapters for the *Handbook*, if the earlier Murchison volumes are taken as a basis of reference, and their inclusion reflects the importance which became attached to the area of applied experimental psychology during the war years.

The research described in these chapters has as its objective the improvement of some aspect of the performance of workers at their jobs. We generally think of four broad approaches to such improvement. These involve changes in personnel selection, changes in training, changes in equipment design, and changes in work methods. The sequence of chapters, as may be noted from their titles above, follows this natural and conventional breakdown. In general the distribution of discussion by topics and subtopics is good, but some allocations such as the following are surprising: (1) The chapter on equipment design is as long as those on the older and more familiar topics of selection and training combined. (2) The criterion problem, which is common to research presented in all four chapters, receives relatively little attention. Bechtoldt and Wolfle each treat the problem very briefly, but in conveniently supplementary ways. Fitts does not discuss it directly, but suggests its importance as he spells out the conditions under which given results may be expected to apply. (3) Reaction time receives essentially no discussion. It has but three entries in the index. (4) The subject of work decrement is dealt with very briefly, in marked contrast to its treatment in Robinson's chapter in the 1934 *Handbook*. There is no disputing that this change is in the right direction, but it is to be regretted that Seashore completely overlooked the recent papers on work under special environmental conditions as well as those on alertness and vigilance over long work periods.

Bechtoldt's chapter on selection is excellently done. He writes only of methodology, of selection procedure, and the statistical aspects of a selection program. He makes no reference to the utility of particular tests in selecting personnel for given jobs. Expressing the caution that each new selection problem "has a few elements in common with old ones, plus a number of novel aspects, the effects of which can rarely be accurately predicted," Bechtoldt refrains from generalizing on test applications, and concentrates instead on methodology as the only strictly generalizable aspect of research in the selection field.

Wolfle, on the other hand, discusses research-established generalizations about learning and considers their application in training programs. The learning principles which Wolfle singles out for consideration and substantiation are six in number and are not exactly homogeneous. Four of them are capable of more or less direct application; somewhat paraphrased, they run as follows: (1) give knowledge of results in as precise a form as possible and as soon as possible after each trial; (2) minimize the possibility of habit interference; (3) vary

practice materials in the interests of more generalized learning; (4) teach "principles" in the interest of generalization. Wolfle's two other training principles imply research needs. They point up the importance of differences in training methods and of the amount of direct guidance given during training. But here research is needed to determine optimum conditions for each training program. Wolfle closes with a short section on the problems met when conducting such research.

Fitts in the chapter on "Engineering Psychology" discloses that this is an area of research where broad generalizations are yet to be formulated. Most of the experiments answer more or less specific questions about the design of particular pieces of equipment. Fitts covers a great many of these studies relating to the design of equipment displays and controls. The implication is that since most of these data have been obtained for equipment components, and have been based upon explicitly stated speed or accuracy measures, the results should be applicable in essentially all work situations where those components are to be used with a comparable emphasis on speed or accuracy. In only a few cases is it known to what extent design recommendations may be generalized to markedly different displays or controls or to operating situations where the requirements are different. Noteworthy in Fitts's chapter is the section on the analysis of motor responses and his examination of the data on the "linearity" of the human system. These are data that have not been brought together before.

Seashore's discussion of "Work and Motor Performance" is directed primarily to the development of the hypothesis that individual differences in performance at jobs involving skilled action arise primarily from differences in work methods rather than from differences in measurable motor skills. He bases his conclusion mainly upon research from his own laboratory, upon factor analysis data which show motor skills tests to involve numerous, narrow group factors, and upon the generally reported low validity of motor tests for predicting job success. In this presentation, relatively little consideration is given to the extensive work of the Army Air Forces Psychology Program on apparatus tests for aircrew selection and no reference is made to Melton's volume in review of this work.

In edited volumes of the size (and weight) of the *Handbook*, individual authors are usually permitted to employ terminologies of their own preference. The chapters under consideration here present no special terminological problems, but it is of some interest to note that whereas Stevens would liberalize the use of the term "measurement" (see Chapter 1), Bechtoldt is strongly conservative and speaks only of "classification" or of "placing people into categories."

BOOK REVIEWS

DOLLARD, JOHN, & MILLER, NEAL E. *Personality and psychotherapy.* New York: McGraw-Hill, 1950. Pp. xiii+488. $5.00.

This is the first bridge between learning theory and psychotherapy sound enough not to topple before being made into a book. Bridgelike, it is easily distinguishable from the solid bodies of facts and observations it connects. Those who know learning theory will profit from the discussion of psychotherapy; those who practice therapy can learn from the consideration of learning; and those conversant with either field will have questions. This is not surprising, in view of the ambitiousness with which this book undertakes its difficult mission.

The authors decided that before formulating a statement about psychotherapy in terms of the concepts of behavior theory and culture it was first necessary to develop a theory of neurosis and of the role of higher mental processes in solving emotional problems. They have sketched out such a system in convincing detail. However, despite the title of the book, they did not associate their notions with psychotherapy in general. Rather, they produced a book limited to the special case of prolonged psychoanalysis. Freudian theory and Freudian descriptions of therapeutic situations have at times been substituted for actual observations. Such a formulation leaves little room for other schools of treatment, unless it explains their effectiveness in psychoanalytic terms or dismisses them as ineffectual. The need for long-term therapy and the importance of transference are astutely deduced from principles of learning, but the coincidental similarity between these conclusions and those of Freud is so striking that one cannot help but suspect a *tour de force.*

The repeated inductions from animal to human learning will cause adverse criticism, especially from clinicians. Although the authors have a well-thought-out position on this complex issue, they do not give it enough emphasis to keep it clearly in the minds of all readers, particularly those outside the psychological fields whom they have stated they mean to include in their audience. Frequently, findings of animal experimentation are applied without apology to psychotherapy with humans. Of course scientists can do things with animals they cannot do with people, but a great deal more can and should be done experimentally with organisms higher than rats before such vast generalizations are made to human beings. This sort of induction would seem more acceptable if experiments were conducted systematically up the phylogenetic scale until it was clearly shown that a principle of behavior remains constant in animals of many sorts, regardless of changes in neural and endocrine structure.

In discussing learning, much attention is given to secondary reinforcement and learned drives. Fear, according to the authors, has certain properties of primary or innate drive. It directly involves reduction of a strong stimulus, and further learning, it is demonstrated, can occur as a result of fear-reduction. Other learned drives are discussed in the same vein, and Wolfe's and Cowles' experiments on token rewards using chimpanzees are cited as examples of the acquisition of the learned drive-value of a stimulus. Yet it is obvious that behavior evoked by a secondary drive like fear is quite different from that motivated by secondary drives which do not involve fear. Secondary drives based on hunger, for example, seem always to involve increase of the strength of a stimulus rather than reduction of a stimulus. Chimpanzees will hoard tokens, handle them, put them in their mouths, and so forth. This distinction is important in view of the authors' definition of reward in terms of a sudden drop in the strength of a stimulus. Consequently, an adequate explanation of an important class of behaviors that involve, for example, reinforcement from earning money, is not forthcoming. Although the reinforcement value of tokens depends ultimately upon primary reinforcement, this fact does not explain the differences in behaviors motivated by such learned drives as fear and money.

Instead of stating that fear is a definite stimulation or pattern of stimulation, the authors use a functional definition and "call anything a stimulus that seems to have the functional properties of a stimulus and anything a response that seems to have the functional properties of a response." This seems innocuous until we turn to the discussion of the higher mental processes and discover that thoughts and images are described as cue-producing responses. Presumably thoughts, and almost certainly images, may originate centrally. Thus what has been of much concern under the heading of cognition and the cognitive theory of learning is covered here by consideration of thoughts and images as response-produced cues that obey the laws of external responses and and that produce the same functional properties as do external stimuli. One may question whether cognitive problems can be disposed of so easily. Though the subtitle of the book says it deals with the topic in terms of learning, thought, and culture, much of the research on thinking is neglected. Of course no book can contain everything—even all the relevant facts on learning do not appear here. The emphasis on "higher mental processes," incidentally, gives an unfortunate impression that the "therapeutic work" of the patient is primarily intellectual rather than affective. Certain passages counteract this impression, but not so strongly as might be desired.

Throughout the section on psychotherapy the reasoning is loose and the approach is almost popular, no matter how technical the terminology. The notions of "suppression," "repression," and "uncon-

scious processes" are handled with a brevity not commensurate with the authors' lengthy study of these matters. The behavioristic notion that conscious mental processes are merely the "saying of sentences" and that in repression these sentences are not said harks back to peripheral or sublaryngeal theories of consciousness that tread roughly on the epistemological problems involved. It is clear enough, though, that the book was not written by J. B. Watson.

A commendable attempt at operational definition is made, but unfortunately much of the subtlety of psychodynamics is lost in this particular translation. The authors did not encompass many phenomena that psychoanalysts have observed, and so leave themselves open to the charge of superficiality. This is particularly true of the discussions of transference and of therapeutic techniques. Perhaps in the future they will attempt to fill in for us more of these details.

The role of the therapist in evoking transference behavior is barely considered. No mention is made of the relationship of the initial behavior of the therapist in defining his role and that of the patient to (*a*) the evocation of transference reactions, (*b*) the continual change of role of the therapist as therapy progresses, or (*c*) the fact that transference reactions in the analytic situation tend to follow a fairly well-defined sequence which Freudian writings imply are coordinated with changes in behavior on the part of the therapist.

There are many strong points in this book. The authors are expertly qualified to expound on the topic of learning and they present good translations of certain neurotic mechanisms into Hullian terms. From intimate acquaintance they discuss Freudian theory and the chief principles of good psychoanalytic practice. It is a thoroughly stimulating and provocative pioneering attempt to close the gap between behavior theory and psychotherapy, connecting two domains with greatly divergent standards of precision and valid evidence. If in the process of connecting the two some of the order in each is lost, the degree of success nevertheless justifies temporary confusions which the authors' next work may help to dispel.

JAMES G. MILLER
JOHN M. BUTLER.

University of Chicago.

MILLER, J. G. (Ed.). *Experiments in social process: A symposium on social psychology.* New York: McGraw-Hill, 1950. Pp. viii+205. $3.00.

The reader's expectations will determine, in large part, his reactions to this book. If he expects to find that the book, as the blurb states, "is directly applicable to contemporary issues," he is apt to be disappointed. If, because of the format (standard McGraw-Hill Publications

in Psychology) and title, he expects to find a careful analysis of different research methods in social psychology or a compendium of recent research in social psychology, he will be disappointed. If, on the other hand, he comes to "hear" a group of workers in social psychology raise and discuss a number of important problems, he will be highly gratified.

The book is a report of a symposium held at the University of Chicago in 1947. That year, it may be remembered, was characterized by the beginnings of the current international political alignment and by fear of atomic warfare. With the tense political situation as immediate background, the symposium was held "to form some idea of the present status of the field of social psychology and to compare and exchange ideas and techniques" and, implicitly, to consider in a preliminary fashion possible contributions of social psychology to socio-political affairs. The symposium consisted of a set of eight papers and a round-table discussion by the members of the symposium and Leo Szilard, biophysicist at the University of Chicago. A central theme runs through all of the papers: If the problems of human interrelationships are to be solved satisfactorily, there must exist a *scientific* basis for the solutions to the problems. Also, there is a recurrent emphasis upon the importance of empirically supported theory and upon the necessity for cross-checking information obtained from laboratory and "real-life" conditions. The papers serve as illustrations of the *kinds* of problems which may have systematic and/or applied value, and of the *kinds* of methods which may prove fruitful. There is no attempt to state all of the important problems, nor is there an attempt to list all of the methods which might be used. The range of problems and methods may be evaluated from a brief listing of topics:

One paper (L. Festinger) deals with laboratory experiments on the effects of group belongingness on voting behavior. Two papers report research using survey methods: one (D. Cartwright) involves estimates of postwar redemption of war-bonds; the other (D. Katz) deals with the evaluation of morale of Germans as related to intensity of bombing of Germany, and of worker morale as related to in-plant and out-plant conditions. A fourth paper (J.R.P. French, Jr.) presents a series of field experiments which are designed to investigate the effect of leadership techniques and group structure on factory output. The last research paper (D. V. McGranahan and I. Wayne) describes a comparative study of German and American characteristics as obtained from a content analysis of successful plays in both countries. These research papers are preceded by a paper which presents a job-analysis of research activities (D. Marquis) and one (R. Lippitt) which emphasizes and illustrates the interactions among field survey, field experiment, laboratory experiment, and clinical analysis of the single case (i.e., the diagnosis which leads to the application of scientific principles). They are followed by an evaluation of problems in social psychology which should, and according to the author (J. Gibson) must, be coped with by learning theory. The round-table discussion of "Social Psychology and the Atomic Bomb" closes the book.

The specific content of the papers, although generally very good, is less important than the impact which the symposium-as-a-whole makes. In part, this is accounted for by the fact that much of the material has been published since the time of the symposium and will be familiar to many readers. More important, however, is the fact that the participants present their material in a manner which is designed to be thought-provoking. In this attempt they are very successful. Whether they talk of the problems of making UN committees more effective, or of the best ways of using survey material, or of the conditions under which field or laboratory experiments should be conducted, the discussion always points to a number of important methodological, conceptual, or applied problems.

Much the same pressures exist today, perhaps with greater intensity than in 1947, for the social psychologist to accelerate the development of his field. In order to increase the rate of progress, it seems very worth while to hold symposia which "take stock" of advances in the field and which use live social issues as a basic setting. The reviewer would like to see this kind of symposium held periodically and to have the reports published with a much shorter time-lag than three years.

ALFRED F. GLIXMAN.
The Training School at Vineland, New Jersey.

THOMPSON, LAURA. *Culture in crisis.* New York: Harper, 1950. Pp. xxiv+221. $4.00.

This is a report of an interdisciplinary research project intended to measure the effects of current Indian Service policy on the Hopi Indians, and to offer a basis for advising the Indian Service how to help the Hopi deal with a severe economic and social crisis which threatens the survival of their tribal society. In 1933 the Indian Service changed from trying to assimilate the Indians into the general American population to a policy fostering the survival of Indian societies through maintaining their economic resources, granting self-government, and encouraging cultural autonomy.

The Hopi project was part of the Indian Personality and Administration Research which included projects with the Navaho, Papago, Sioux, and Zuni. The Research was initiated by the United States Office of Indian Affairs. Laura Thompson, an anthropologist, coordinated the Hopi project which was conducted by a large staff of anthropologists, sociologists, psychologists, psychiatrists, educators, and other specialists who served as research planners, trainers of field staff, field workers, and data analysts. The project exemplifies best practice in action research since Indian Service administrators and leaders in the communities studied were involved whenever possible in planning and conducting the research and in evaluating and applying

the findings. The concluding chapter of *Culture in Crisis* offers suggestions to the Indian Service on dealing with the Hopi crisis.

While the project was undertaken to provide information useful to administrators, the study has significance for the theory and methdology of basic research in culture and personality. The researchers were prevented from doing a straightforward before-and-after study since the new policy had been in effect for several years when the project was begun in 1941. In the effort to establish what Hopi society was like before 1933, recourse was had to historical sources and to previous anthropological research. Two Hopi communities known to differ considerably in present social structure were selected for study. The attempt was made to "reconstruct their common cultural heritage" as a way of approximating the basic Hopi culture which both communities had shared before economic and social factors caused them to diverge. This method of historical reconstruction, despite its methodological limitations, appears to have been of great value in directing the researchers' attention to factors important for understanding differences in Hopi communities and differences in the personalities of their members.

In the research a major emphasis was placed on the "needs, trends and values" of Hopi individuals. Hopi society was evaluated in terms of how it affected the "biopsychological health" of individuals. Research on Hopi personality was focused in a psychological appraisal of 190 Hopi children from the two communities being studied. Intelligence was measured by standard performance tests, the Rorschach and TAT were administered, and several tests designed to measure personal values and ideology were given. Although these data were analyzed statistically, the findings are not reported in a form which permits the reader to judge their significance. Indeed, throughout the book, the reader who wishes a look at the data on which generalizations are based is repeatedly frustrated, though footnotes refer to other and more detailed reports of the research.

Rorschach interpretation (Klopfer) is presented uncritically, as are other "depth psychology soundings." Hopi personality is described as "unusually complex," "unusually balanced" with a "nice adjustment between the expression and control of psychic forces." The Hopi are said to "approach problems as organized wholes." Obviously, this sort of personality analysis leaves much to be desired.

To the student of culture and personality, *Culture in Crisis* offers a provocative analysis of differences between the two Hopi communities studied. In one of them, the traditional Hopi way of life had been seriously disrupted by social pressures from outside. Evidence is offered to support the view that the conflict of Hopi and non-Hopi cultural patterns and ideologies within this community is reflected in the personalities of its members.

The book contains a stimulating analysis of the Hopi "symbol system," including an excellent chapter on the Hopi language condensed from the writings of the linguist, Benjamin Lee Whorf. This discussion makes a good case for the view that Hopi language and ideology play a vital part in determining the individual's self concept and his adjustive behavior.

Dr. Thompson places great stress on a multidimensional approach which synthesizes "ecologic, somatic, sociologic, psychologic, and symbolic" factors. She presents a theoretical orientation developed by the staff of the Indian Personality Research which is intended to comprehend the interacting roles of these factors in the "dynamics of culture structure." It would be too much to expect or hope that this approach would live up to the author's claim for it as being "an integrative, psychiatric type of methodology for investigating the communal personality structure in total context" (14). The important point is that the Hopi project, and the projects with the four other tribes, represent an approach to social science research in which members of related disciplines attempt to bring their distinctive theories and methods together in solving research problems. Viewed in this light, *Culture in Crisis* is an important addition to the literature of interdisciplinary research and deserves a place in the company of such books as Leighton's *Governing of Men*.

Culture in Crisis is rich in facts and ideas, is well organized and well written. It should be very favorably received by the student or researcher interested in culture and personality, and by anyone interested in applying social science knowledge to the solution of practical problems.

GLEN HEATHERS.

Fels Research Institute.

KELLER, FRED S., & SCHOENFELD, WILLIAM N. *Principles of psychology: A systematic text in the science of behavior*. New York: Appleton-Century-Crofts, 1950. Pp. xv+431. $4.00.

Although *Principles of Psychology* was written for use in elementary courses, its authors have departed rather widely from the usual practices of introductory textbook writers. Keller and Schoenfeld have not attempted to please all of the potential revenue-yielding consumers of such texts by reviewing all areas and viewpoints in psychology. Rather, they have made a sincere and careful effort to integrate a number of selected topics within a single conceptual framework—that provided by the work of B. F. Skinner. That they have been successful in this attempt will be apparent to readers familiar with Skinner's *The Behavior of Organisms*. Such readers will be struck by the extent to which *Principles of Psychology* constitutes a popularization of the attitudes, methods, terms, experimental findings, and organizational structure of

the earlier work. The parallelism is so close that some critics may regard the title as inappropriate and may favor the substitution of, say, *Skinner for Beginners* as being both more accurate and more euphonious.

The central theme pervading the entire book is simply this: much, if not all, of human and animal behavior is learned; and the learning process in each case is governed by certain "well-established psychological principles." As a glance at the chapter headings will show, the "principles" upon which reliance is placed are terms describing both procedures and concepts in the area of conditioning. Included are such words as respondent and operant conditioning, extinction, reconditioning, generalization, chaining, and secondary reinforcement. The task of presenting, elaborating, interrelating, and applying these "principles" constitutes the major burden of the textual material.

The opening chapter, *Psychology and the Reflex*, is an abridged version of Skinner's first chapter. Psychology is defined as the behavior of organisms; the reflex is distinguished from the reflex arc; Skinner's "static laws of the reflex" are introduced—though not as "laws"; and the notion of reflex strength is propounded. In the next two chapters, the reader is introduced to the concepts of respondent and operant conditioning, along with rudiments of the corresponding experimental methods. Differences between Type *S* and Type *R* conditioning are stressed, *elicited* behavior and *emitted* behavior are contrasted, and the groundwork is laid for subsequent interpretations of human behavior in terms of variables significant for operant conditioning.

Chapters 4 through 7 treat of extinction and periodic reconditioning, generalization and discrimination, response variability and differentiation, and chaining. At this time, the reader may first become aware of the considerable skill with which the authors have enriched each chapter by the introduction of some of the more traditional topics of psychology. For instance, in the generalization-discrimination chapter, one finds, in addition to the expected chapter-heading material, some treatment (though usually cursory) of each of the following: similarity, psychophysical methods, experimental neurosis, simple and complex reaction time, depth perception, concept formation, and transposition. And again, in the chapter on chaining, short sections on each of these topics have been adroitly interspersed: proprioception, thinking as inner speech, sensorial and muscular reaction times, the context theory of meaning, word association experiments, guilt detection and complex-indication, maze learning, and serial verbal learning.

The remaining chapters, on secondary reinforcement, motivation, emotion, and social behavior, continue strongly to reflect both Skinner's concepts and his data. To an important degree, however, these and many other chapters are liberally sprinkled with experimental findings from Columbia psychologists and from other workers; notably, Hull and his students, Thorndike, Muenzinger, Mowrer, and Estes. Unfortunately, unpublished Columbia studies and even "personal communications" are occasionally cited in support of debatable issues in lieu of already-published experiments of perhaps equal merit and relevance.

The opening sections of the concluding chapter (Social Behavior) deal with the manner in which the cultural environment is presumed to mould individuals and groups. At each point, the moulding process is to be understood as involv-

ing the processes of reinforcement and extinction, the establishing of discriminative stimuli, the differentiating out of responses, the formation of response chains, and the like. The work concludes, having come full circle, with an abstract of more of Skinner's work; in this case, his treatment of language as given in his *William James Lectures on Verbal Behavior* at Harvard.

The task of attempting wisely to criticize a work such as this is made exceedingly difficult by the marked extent to which it represents a rephrasing of Skinnerian psychology. If the critic detects a weakness, at whom should the accusing finger be pointed? At Skinner or at Keller and Schoenfeld?

Probably the most frequent general objection that will be raised concerns the failure of the text to provide the beginning student with material calculated to aid him in the development of a "scientific superego." Little or no stress is placed upon the need for the careful design and rigorous control of experiments; statistics is apparently never mentioned, much less its use in the testing of hypotheses; and though a separate section is devoted to the treatment of intraindividual variability, nothing is said about interindividual variability. Such material could, of course, be introduced into an elementary course by way of supplementary lectures or in the laboratory. But students who confine themselves largely to the text might well acquire the conviction that any difference is a *significant* difference; even differences whose detection requires the foreshortening of cumulative curves by holding the page parallel to the line of sight.

Although the authors have broadened Skinner's assembly of descriptive concepts by incorporating additional experimental and illustrative material, they have retained a number of his more debatable, less elegant notions. The term *reflex*, for example, is still awkwardly employed both to denote objective stimulus-response relations and "... to denote responses for which related stimuli are not clearly observable" (p. 4). This usage stems directly from Skinner. He had insisted initially, it may be recalled, that the use of the word *reflex* was to be limited strictly to correlations between observable stimuli and observable responses. But he immediately applied the term to behavior (operants) whose distinguishing characteristic was the presumed absence of observable stimuli. Moreover, in stressing the correlational nature of the reflex he had held that it must be "... emptied of any connotation of the active 'push' of the stimulus" (*Behavior of Organisms*, p. 21). Yet it is hard to imagine a more vigorous metaphorical "push" than that conveyed by his terms *elicited* behavior and *eliciting* stimuli. Skinner himself has perhaps never taken this matter too seriously, since he has seemingly not attempted to formulate precise criteria for deciding, in any given case, whether a response is elicited or emitted. Needless to say, psychologists do not yet agree that this distinction can be made in a clear-cut fashion or even that it is useful to make the attempt.

It seems safe to surmise that Keller and Schoenfeld were on some occasions torn between the desire to retain Skinner's usages intact and the desire to modify them. As a case in point, it is asserted (p. 50) that neither latency nor response magnitude is a satisfactory measure for determining the strength of an operant. But a number of experimental studies, e.g., those of Graham and Gagné, and of Zeaman, in which starting-time was used as a measure, are cited with complete approval. Moreover, rate of responding is said to be the best measure of operant behavior, at least in the early chapters. But it is essentially abandoned, as probably it must be, whenever an attempt is made to apply the "principles" to adult human (operant) behavior.

In their preface and elsewhere, the authors declare their intent to present an integrated treatment of selected psychological topics. The systematic framework to be used in achieving this integration is termed *reinforcement theory*. This is rather puzzling, of course, because the Skinnerian system of observed regularities in behavior has seldom been dubbed a "reinforcement theory." Skinner would certainly qualify as a *general* reinforcement theorist in that the growth of reflex strength is assumed to depend on the presentation of (empirically catalogued) reinforcing stimuli. But he is not a drive-reduction reinforcement theorist. It is to the latter that the term *reinforcement theorist* has most usually been applied.

The authors' reference to the integrative function of theory is also of interest in the light of Skinner's recently expressed doubts as to the utility of learning theories. He has always apparently held that theorizing is fun though hardly worth while and that psychology should restrict itself to the description of behavior and to the determination of the variables of which it is a function. Actually, in Keller and Schoenfeld's book, the *structure* of an integrating theory is never in evidence, though a fair degree of integration has nevertheless been achieved. It would appear, therefore, that the integration stems not from theory, but from the consistent, and often highly insightful, application of such terms as acquisition, extinction, partial reinforcement, and secondary reinforcement to examples of everyday human behavior.

One of the important positive contributions of the text lies in its having achieved at least a partial coalition of Skinner's data and terms with the data of other behavior scientists. Certainly any attempt to promote unity among superficially different "schools" must be applauded. Perhaps it is not too much to hope that Keller and Schoenfeld's example will be followed by others and that the day will be hastened when psychologists will use the same terms when talking about the same things.

With the possible exception of the somewhat mawkish pontifications of the concluding page, the style in which the material is presented is unquestionably superior. The authors have an enviable ability to

write with simplicity and clarity while retaining their adult, scientific integrity. Doubtless they have heard of the Flesch count. But they have not been led thereby into using the prattling, first-reader-like sentences one finds in a few recent texts. They seem to have striven to make the text eminent *and* readable, not merely "eminently readable."

As used at Columbia, where it is accompanied by carefully planned bar-pressing experiments, *Principles of Psychology* is said to be highly successful, at least in arousing student enthusiasm. Whether it will prove equally stimulating at institutions devoid of comparable laboratories, and whether, in addition to arousing interest, it will provide the novice with training of adequate breadth and rigor, are questions that must be left to the future.

JUDSON S. BROWN.

State University of Iowa.

SPEARMAN, C., & JONES, L. WYNN. *Human ability.* London: Macmillan, 1950. Pp. vii+198. $2.50.

The slender physical dimensions of this volume conceal what has obviously required enormous amounts of mental and experimental effort, over the course of many years. Certainly, those whose work is reported within its covers might desire a more weighty memorial; however, size alone is no index of scientific importance, and in the present case the small bulk should serve to promote awareness and reading by a wider audience. The epigrammatic style of the presentation has produced many easy-to-quote passages, and there is little doubt that much capital will be made of the fact. Perhaps the authors' opening statement of the "new 'menace' " posed by the widespread utilization of mental tests will receive the widest currency.

The reviewer would hesitate to apply the term scholarly to a work with so little depth of focus, i.e., with so little apparent positive effort to integrate disparate results into a coherent treatment of ability. It is perhaps unfortunate that Dr. Jones, in preparing Spearman's manuscript for publication, was not less reluctant to expand and develop the topics more fully. The reader is left with the impression that the generalized theory of two factors, previously expounded in the *Abilities of Man* and elsewhere, is too inflexible to account for all the results which are now in. It has not been shown, for example, why a determinate simple structure can repeatedly be found in any sufficiently large battery of ability measures. Such a phenomenon is just as real as a correlation hierarchy. That a resolution of the position of the two-factor and multiple-factor theories is possible may be seen in some of Guttman's recent work. In other instances the theoretical colorations of the presentation are even more evident. Thus, Thurstone is strongly criticized for rotating his factorial axes to simple structure

in a particular study while Alexander, a few pages later, is presented favorably despite his use of essentially the same methods.

The principal virtue of the present volume is its combination of breadth of scope, foreshortened style and singleness of viewpoint. *Human Ability* does present an up-to-date summary of many of the results which have bearing on the half-dozen or so factors which have fallen heir to the original theory of two factors, together with some further elaborations and refinements of the theory. There now seem to be three general factors—Abstract neogenesis (*G*), Perseveration (*p*) and Oscillation (*O*)—at least four broad group factors, including verbal, mechanical, fluency, and memory factors, and the ubiquitous specific factors. There is much in this analysis of human ability to suggest important areas for further research, and it would seem especially appropriate for American psychologists to give this volume thoughtful consideration.

D. R. SAUNDERS.

Educational Testing Service.

LINDQUIST, E. F. (Ed.) *Educational measurement.* Washington, D. C.: American Council on Education, 1951. Pp. xix+819. $6.00.

According to the preface, the purpose of this volume is to provide a "comprehensive handbook and textbook on the theory and technique of educational measurement." In addition the preface points out that this volume was inspired by the fact that there existed no suitable reference work that could be used in advanced courses in educational measurement at the graduate level.

It is extremely difficult to review a work so encyclopedic in character and so heterogeneous in authorship—because of the varying quality of different parts of the work, and because different authors have different audiences in mind and hence write at different levels of difficulty. The difficulty is enhanced in the present volume, where chapters vary from the platitudinous to the profound. Of course, what is a platitude to one still may be a profundity to another. The part of the volume which seems to come closest to the stated purposes is Part 3 on "Measurement Theory." The five authors of this section reflect in their writing a level of scholarship far above that found in most of the rest of the volume. The chapter by Irving Lorge on "The Fundamental Nature of Measurement" is particularly notable and represents a real contribution to the understanding of the philosophical problems underlying psychological measurements. Robert Thorndike's chapter on "Reliability" provides a well-organized account of the topic. Edward Cureton's chapter on "Validity" is valuable reading matter for the professional psychologist. A fourth major contribution is made by John C. Flanagan in his chapter on "Units, Scores and Norms," which is one of the best accounts the reviewer has yet seen concerning the

development of norms. This chapter, like the three chapters which precede it, provides an organized theoretical framework which is so lacking in most other works on educational measurement. Finally, Charles I. Mosier completes this section with a discussion of problems related to the use of "Batteries and Profiles."

The remainder of the volume presents a much less happy picture. To the authors of the first four chapters on "The Functions of Measurement in Education" must inevitably fall the lot of saying what has been said many times before. They discuss the usual generalities about the functions of measurement in a commonplace way. Often the discussion is vague and general. For example, in the chapter on "The Functions in Measurements in Educational Placement" there is extensive discussion of what is described as the problem of articulation in education. The nature of this problem really is left up to the reader except for such statements as, "The problem of articulation has to do with techniques and procedures for bridging these transitional points so that there is continuity in the educational program and so that each student is enrolled in courses which are consistent with his interests and level of academic proficiency." The educational problem of articulation is never properly identified and all that is said is that here is some kind of problem which perhaps people in educational measurement can clear up.

Part 2 of the book is devoted to a presentation of techniques used in the construction of achievement tests, and treats such varied topics as methods for reproducing tests, item selection techniques, the essay examination, and the planning and development of objective tests. The most notable contribution of this part of the book is a chapter by Frederick B. Davis on "Item Selection Techniques." This chapter is commensurate in scholarship with those in the final part of the volume. Particularly inadequate is the chapter on "Planning the Objective Test." In this chapter there is practically no discussion of the difficult but important problem of defining a domain of behavior within which measurement is to take place and yet this is a fundamental step in the handling of achievement tests. The omission of a proper discussion of how to specify what a test is to measure is not rectified by an adequate discussion of the topic in the chapter on "Writing the Test Item." The latter chapter is devoted almost entirely to the discussion of the form of test items rather than of their function. For example, there is an excellent discussion on suggested rules for writing multiple-choice items and yet one cannot find out the kinds of purposes for which multiple-choice items might be used. The paragraph on the applicability of multiple-choice questions begins with a statement, "The multiple-choice form is widely applicable," and nothing more is said concerning the kinds of achievements which they may be used to measure. Of what avail is a multiple-choice item which illustrates all the suggested

rules of good item writing if it measures an achievement which is trivial or irrelevant? Some contributors to this volume might well be reminded of the dictum of a famous architect stated in an analogous context that "form follows function." The function of a test item must be a determiner of its form, and yet in this volume the various forms of test items are discussed without relating them to the functions they may serve.

Finally, there is one major criticism of the book as a whole which cannot be omitted. The work seems to imply that educational measurement consists mainly of the measurement of academic achievement of the type which has been stressed in traditional schools. It is to be hoped that advanced courses in measurement for which the book is designed will cover some of the broader aspects of the measurement of growth. Educational measurement cannot be preoccupied only with intellectual development in an age when so much stress is being placed on the nonintellectual aspects of growth. Much has been done already to assess the development of interests, values, attitudes, and other aspects of personality and these must surely be given some place in an advanced course in measurement. Unless this is done such courses will be geared to an educational philosophy which is no longer too widely accepted.

ROBERT M. W. TRAVERS.

Teacher Education Division, Board of Higher Education, New York City.